Data Mining Project

Group 2

Audit Fraud Risk

Samuel Chadick, Adam Rodriguez, Samantha F, Rebeca Duron, Jessica Rodriguez, Sachi Wijeratne, and Elizabeth Monks

G. Brint Ryan College of Business, University of North Texas

DSCI 4520: Introduction to Data Mining

Dr. Mahdi Ahmadi

December 15, 2023

# Table of Contents

[Table of Contents 2](#_Toc114622770)

[Abstract 3](#_Toc1224465152)

[Introduction 4](#_Toc735309350)

[Background 5](#_Toc125999031)

[Importance of Solving the Problem 5](#_Toc73092232)

[Mitigating Fraud Risk 5](#_Toc1922872789)

[Optimizing Auditor Resources 5](#_Toc1914809024)

[Enhancing Quality of Audited Work 6](#_Toc721108161)

[Proactive Detection and Prevention 6](#_Toc1241719541)

[Target Variable 6](#_Toc1900818145)

[Results 7](#_Toc2044215613)

[K-NEAREST NEIGHBORS 11](#_Toc1140796218)

[LOGISTIC REGRESSION 11](#_Toc613193623)

[RANDOM FOREST 12](#_Toc82190549)

[Conclusion 13](#_Toc15024353)

[References 14](#_Toc1244567560)

# Abstract

Auditors currently lack an effective tool to classify firms for potential audit fraud based on historical risk factors. The absence of a streamlined process hinders their ability to identify patterns and indicators of fraudulent behavior, compromising the accuracy and efficiency of financial audits. This gap in the auditing process poses a significant challenge, as it jeopardizes the accuracy and reliability of financial reports, potentially allowing fraudulent practices to go undetected. A solution is urgently needed to automate this classification process, enabling auditors to proactively detect and address potentially fraudulent activities, thereby enhancing the reliability of financial audits and overall financial reporting integrity.

# Introduction

The objective is to develop a model to accurately categorize and forecast firms with a higher likelihood of engaging in fraudulent activities. Considering the escalating incidents of financial fraud, the current landscape demands innovative solutions to enhance the efficiency of audit agencies in identifying fraudulent cases. The conventional fieldwork involved in audits necessitates substantial planning, allocation of resources, and time. By developing this model, we aim to streamline and optimize the process, providing a more effective and timely means of identifying firms with elevated fraud risks.

# Background

Auditing has recently been a matter of interest in all business realms. With the introduction of the Sarbanes Oxley Act, established in 2002, all publicly traded companies must be more meticulous with company audits. Audits protect investors from the misrepresentation of financial statements and documents in a company to lure them to invest. Audits happen both internally within a company and externally by third parties who are non-biased. Auditing has also existed for hundreds of years; early traders understood the importance of record-keeping and accuracy. Today auditing is a lot broader and consists of large teams of people, but even today investors are still exposed to fraud due to human mistakes. It would be extremely beneficial for auditors to be able to spend their resources and time on the cases that are more likely to be fraud. Machine learning models can help auditors improve the quality of their fieldwork by predicting firms that are likely to resort to high-risk practices.

Addressing the challenge of fraud risk is paramount due to its far-reaching implications. Minimizing the risk of fraud is not only imperative but also aligns with the strategic goals of audit agencies. Empowering auditors to allocate their resources and time judiciously by focusing on cases with a higher likelihood of fraud is a significant advantage. Machine learning models offer a transformative solution, enhancing the overall quality of audit fieldwork by accurately predicting firms prone to engaging in high-risk practices. This not only fortifies the audit process but also contributes to a more targeted and efficient allocation of resources within the auditing framework.

# Importance of Solving the Problem

## Mitigating Fraud Risk

Addressing the problem is crucial to minimize the risk of fraud in financial reporting. By enhancing auditors' ability to classify firms based on historical risk factors, the overall integrity of financial audits is strengthened, and instances of fraudulent activities can be more effectively identified and mitigated.

Optimizing Auditor Resources

Enabling auditors to focus their resources and time on cases with a higher likelihood of fraud is essential for efficiency. A solution to this problem would empower auditors to prioritize and allocate their efforts more effectively, ensuring that they concentrate on cases that pose a higher risk of fraudulent practices. This optimization contributes to a more streamlined and resource-efficient audit process.

## Enhancing Quality of Audited Work

Implementing machine learning models to predict firms likely to engage in high-risk practices improves the overall quality of audited work. By leveraging advanced analytics, auditors can make data-driven decisions, leading to more accurate and insightful assessments of financial statements. This not only strengthens the credibility of audit reports but also enhances the reliability of financial information for stakeholders.

## Proactive Detection and Prevention

The ability to predict and classify potential fraud cases in advance allows for proactive detection and prevention measures. Rather than reacting to fraudulent activities after they occur, auditors can take preemptive actions to address high-risk situations, reducing the financial and reputational damage associated with fraud.

## Target Variable

The binary variable is classified as the risk flag, with 1 representing audit fraud risk and 0 meaning there is no audit fraud risk present. The predictor is which sectors likely have a higher risk of fraud risk based on their historical patterns. The size of the data set is 776 records. We will be using RapidMiner to build a predictive model for classification. In this model, we are exploring three different algorithms: k-NN, logistic regression, and random forest. For the classification analysis aimed at predicting firms engaged in fraudulent processes, we plan to leverage a combination of data mining techniques and tools. Our approach includes the following;

**Classification Analysis:**

* We will employ machine learning classification algorithms to build predictive models.

**Tools for Data Mining:**

* RapidMiner: We will utilize RapidMiner for its comprehensive suite of data science and machine learning tools.
* Microsoft Excel: Excel will serve as a tool for visualization.

**Exploratory Data Analysis (EDA):**

* Prior to the classification analysis, we will conduct thorough exploratory data analysis (EDA) to understand the characteristics of the dataset.

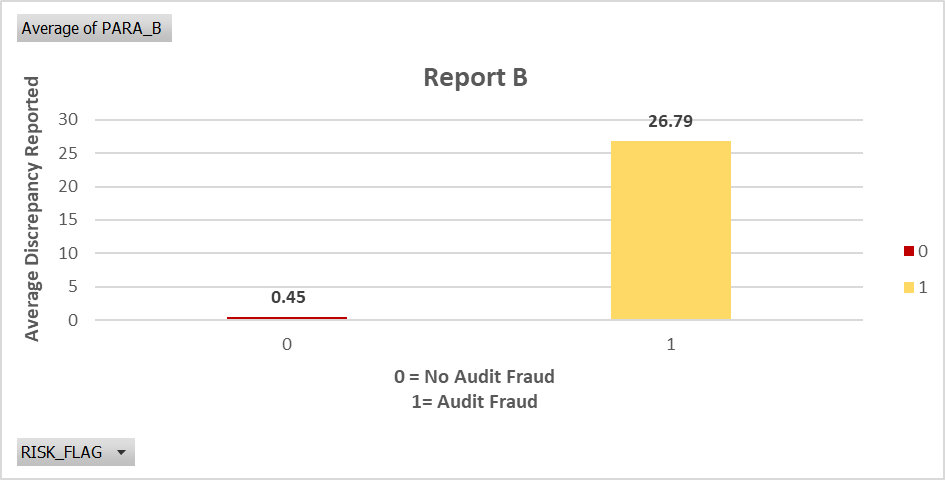
**Data Pre-processing:**

* Data pre-processing steps, such as handling missing values, encoding categorical variables, and scaling features, will be carried out to prepare the data for input into the machine learning models.

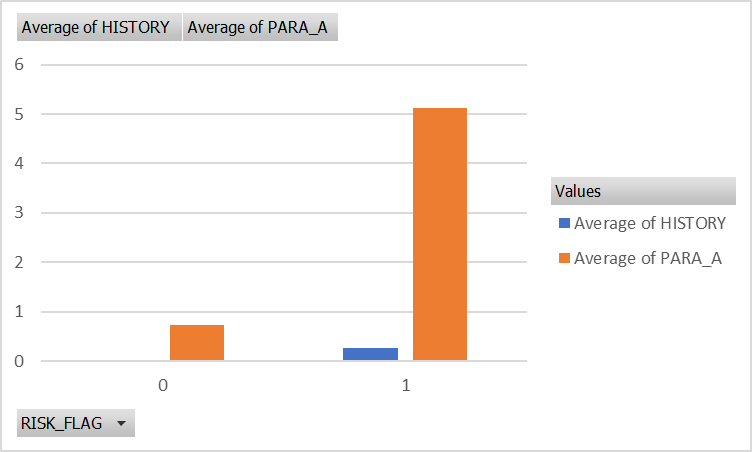
We will evaluate the performance of your model by using Data Exploration and Visualization. We will look at the regression output on excel and summary statistics. Then compare the correlation coefficients, P values, T statics. We can use data visualization to see which firms/sectors have the highest risk of frauds occurring. Using accuracy to also measure our model to determine false predictions. Additionally, we will also use the Confusion Matrix (precision and recall) a ROC Curve (AUC).

# Results

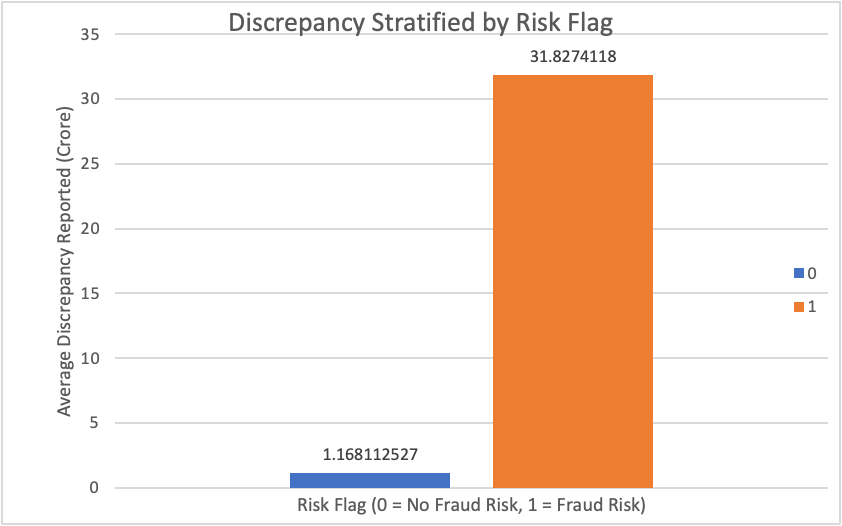
The discrepancy found in the planned expenditure of summary report B in Rs (in crore):



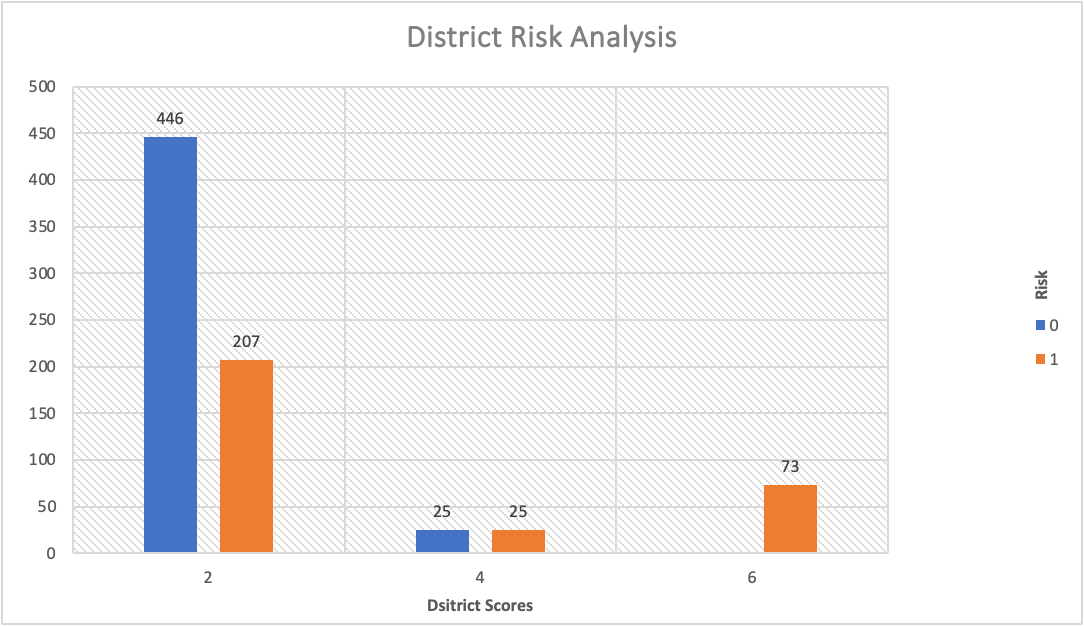
In cases where the audit fraud risk was 0, a discrepancy was found in the planned expenditure of inspection and the summary report B, amounting to an average score of 26.79 crore rupees. Similarly, for cases with an audit fraud risk of 1, a discrepancy was found with an average of 0.45 crore rupees. Based on the insights provided by this model, it can be inferred that in Report B, the level of audit fraud risk was significantly higher.

Behavior of PARA-A and Historical Risk Score of a District for the past 10 years in comparison with risk status class: 

For the pattern in Para A and Para B are similar. Because if the risk flag is 0, both have lower scores. If the risk flag is 1, both have a higher score. This signifies that Para A and Para B have similar discrepancy level for the risk flag.



The above chart measures the average discrepancy total from both PARA\_A and PARA\_B. This data is separated into two categories: records that were flagged as a fraud risk and those that were not. The visualization indicates that the average discrepancy reported for a firm flagged for fraud risk was 30 times higher than a firm that was not flagged.



Based on the provided data for District Risk Analysis, here is an analysis of how historical district risk scores relate to the risk flag (0 or 1):

District Score 2: There are 446 cases with a District Score of 2 and a Risk Flag of 0, while there are 207 cases with a Risk Flag of 1. This suggests that District Score 2 is associated with a higher incidence of audit fraud risk (Risk Flag 1) compared to no audit fraud risk (Risk Flag 0).

District Score 4: In this category, there are 25 cases each for Risk Flags 0 and 1. District Score 4 appears to have an equal distribution between audit fraud risk and no audit fraud risk.

District Score 6: There is a notable difference here, with 73 cases having a District Score of 6 and no information provided for Risk Flag.

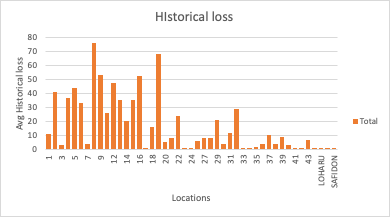
The data suggests that specific district scores may indeed relate to the risk flag, with some districts having a higher likelihood of audit fraud risk (Risk Flag 1), while others have a more balanced distribution or incomplete information.

District Score 2: This district has a "District Score" of 2, indicating a moderate historical risk score. In this district, the "MONEY Value" is ₹8,870.154. This suggests that audits conducted in this district have identified a substantial amount of money involved in misstatements, signifying potential financial discrepancies or issues.

District Score 4:  In contrast, another district with a "District Score" of 4, which is presumably higher in terms of historical risk, has a significantly lower "MONEY Value" of ₹507.45. This may imply that despite a higher historical risk score, the amount of money involved in misstatements in audits conducted in this district is relatively low.

District Score 6: The district with a "District Score" of 6 has a "MONEY Value" of ₹1,579.06. This district falls in between the other two in terms of historical risk scores and the amount of money involved in misstatements.

A sector with a "Sector Score" of 1.85 has a total of 41 risk values associated with it. This indicates a relatively high level of audit risk within this sector, despite the low historical risk score. Another sector with a "Sector Score" of 1.99 has 13 risk values. Again, this suggests a notable audit risk in this sector compared to the sector score. Sectors with historical risk scores of 2.34 and 2.36 have a risk value of 1 each. This signifies a very low count of audit risk flags, indicating a lower audit risk in these sectors. The sector with a "Sector Score" of 2.37 has 36 risk values, indicating a relatively higher audit risk, given its historical risk score. As the sector score increases, we see a notable increase in the Risk values. Sectors with scores like 2.72, 3.41, and 3.89 have 53, 59, and 66 risk values, respectively, suggesting a higher audit risk in these sectors. Moving into significantly higher sector scores, such as 15.56, 17.68, 21.61, 55.57, and 59.85, the counts of "RISK\_FLAG" values remain relatively low, indicating a comparatively lower audit risk in these sectors despite their high historical risk scores.

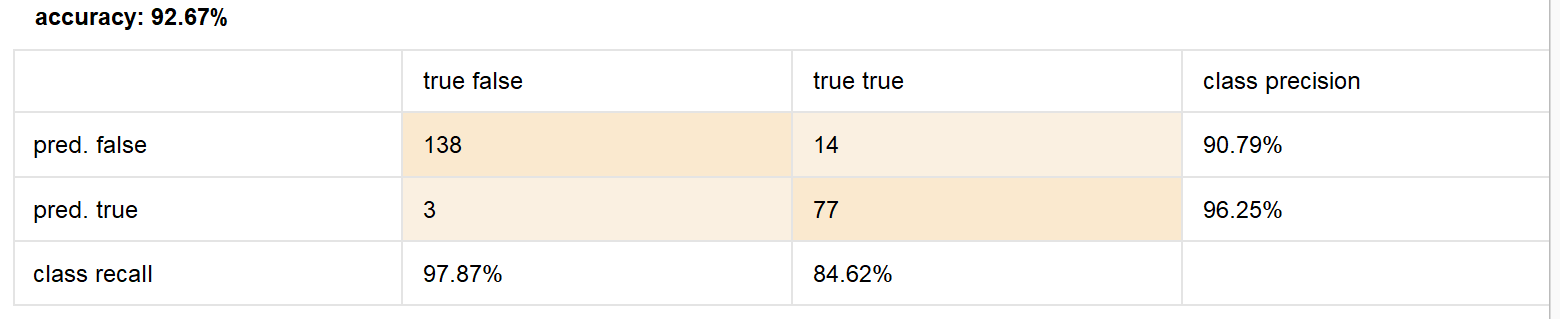


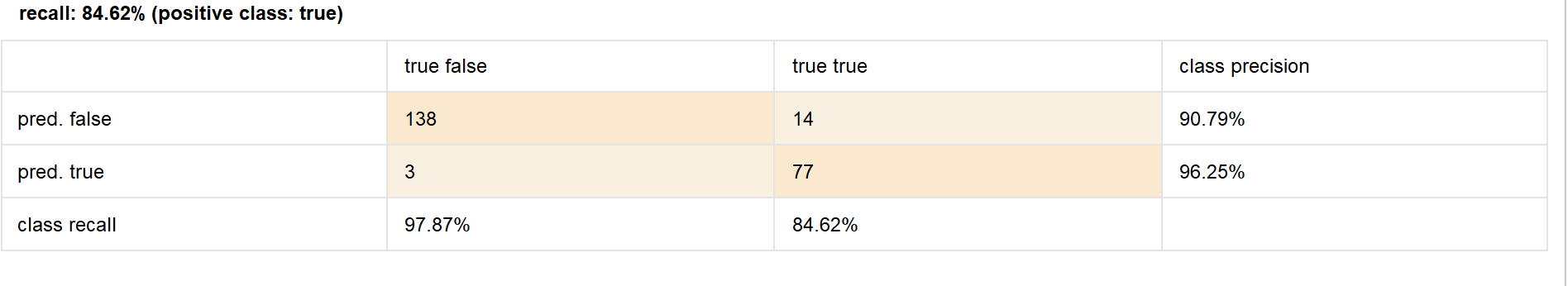
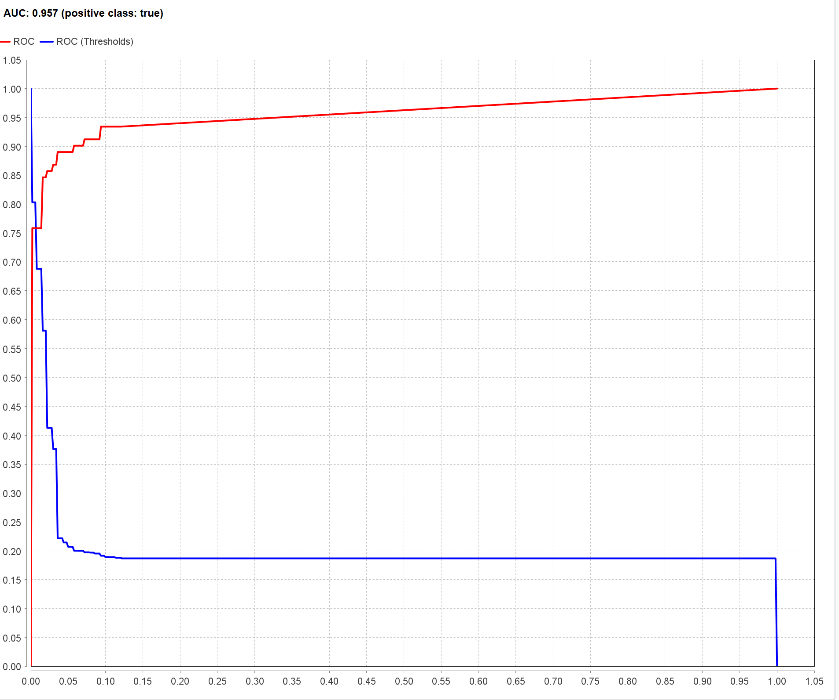
This is the average historical loss per location. We notice that locations 8 and 19 have the highest historical loss. Which means that these locations are more prevalent to future mistreatment of money. Locations 17, 23, 24, 33, 34, 41, 42 have one of the lowest. This means that historically these locations are least prevalent to fraud risk.

Our predictive model is classification by utilizing RapidMiner and exploring k-NN, logistic regression, and random forest.

## K-NEAREST NEIGHBORS

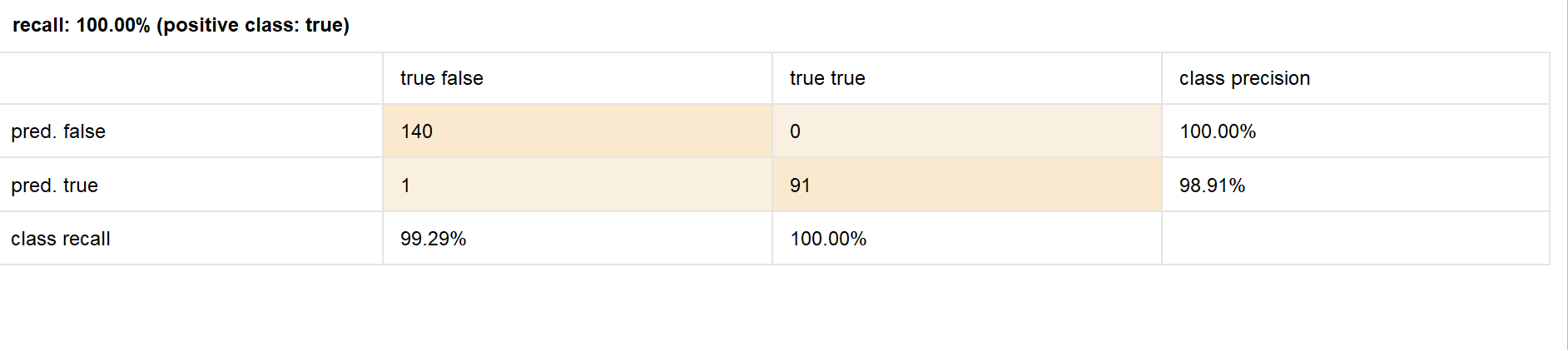
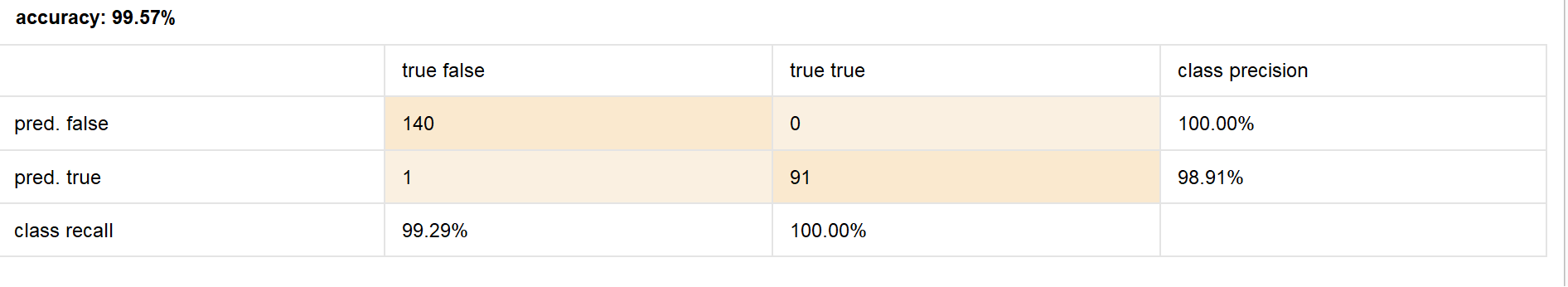
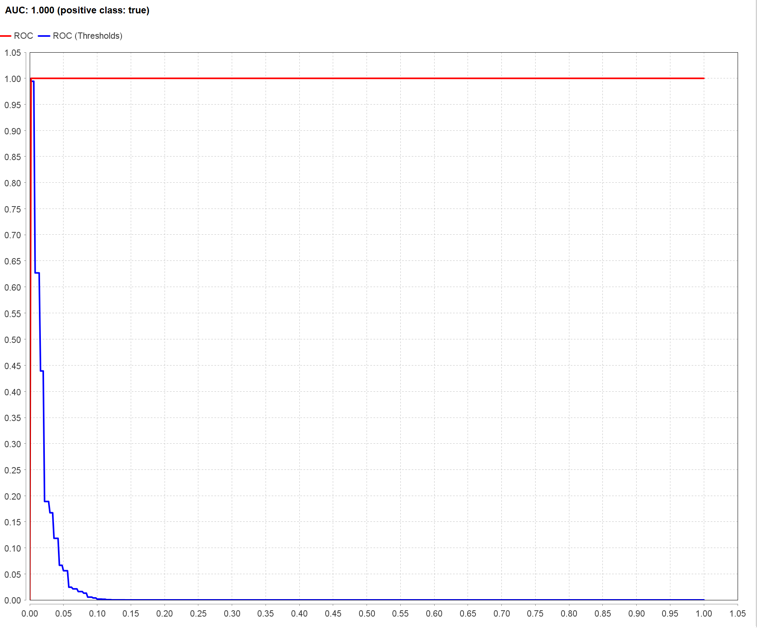
A k-nearest neighbors classification model was chosen as it is useful in predicting a target class (in this case, fraud risk or no fraud risk). For our k-NN classification model, a k-value of 5 was determined as the optimum number of neighbors through a trial-and-error process in which the performance metrics of the model were monitored and compared to one another. The k-NN model was the poorest performing out of the three models created, reporting an accuracy score of 92.87%, a recall score of 84.62%, and an AUC score of 0.957. Performance visualizations for this model are below:





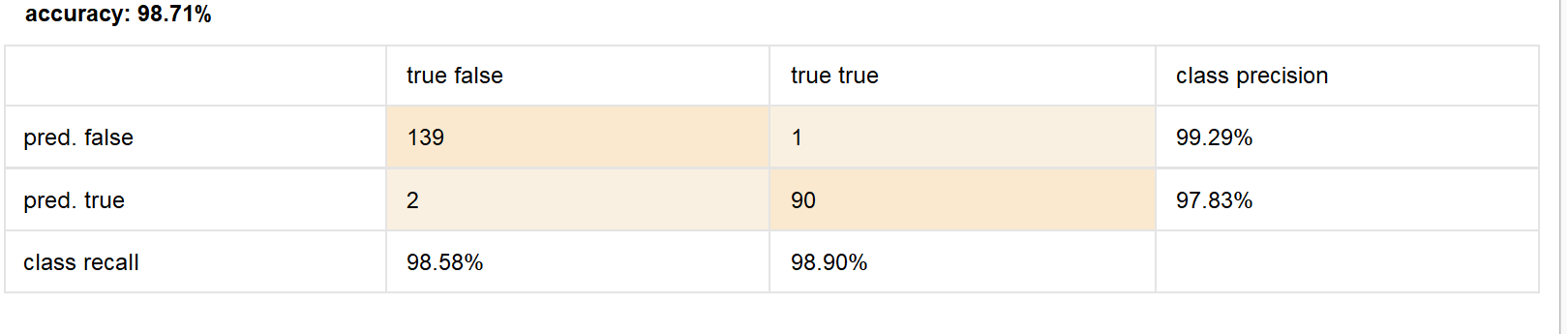
## LOGISTIC REGRESSION

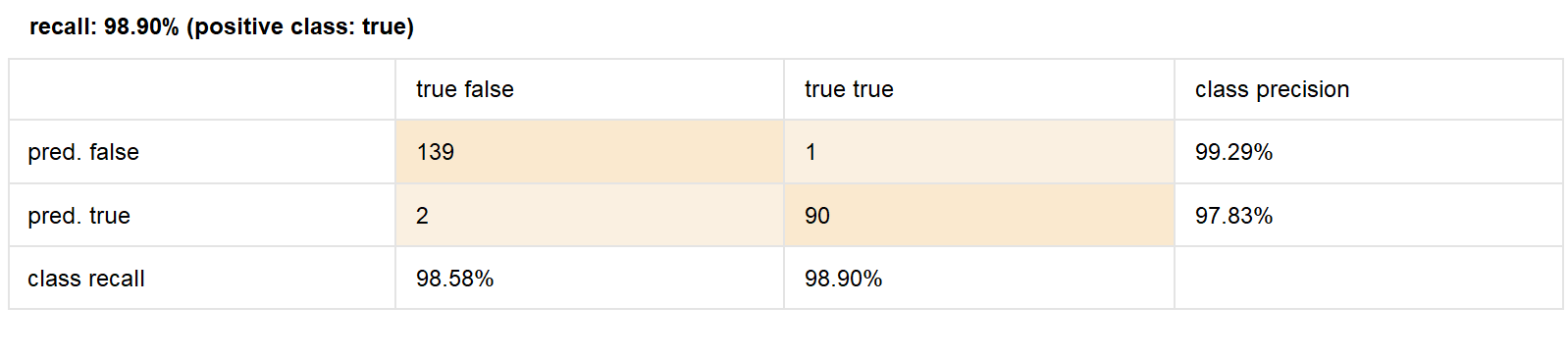
Our second model chosen was a logistic regression model. This model is ideal for predicting a binary target variable and vastly superior in this aspect when compared to using logistic regression. We did not perform any specific tuning of the model, leaving the cutoff score at 0.5, however this model demonstrated superior performance in correctly predicting the fraud risk or no fraud risk classes. The confusion matrix of the model identifies only one error on the validation dataset was made, a false positive for fraud risk. This very low error resulted in high scores across the board in accuracy (99.57%), recall (100.00%) and AUC score (1.000).

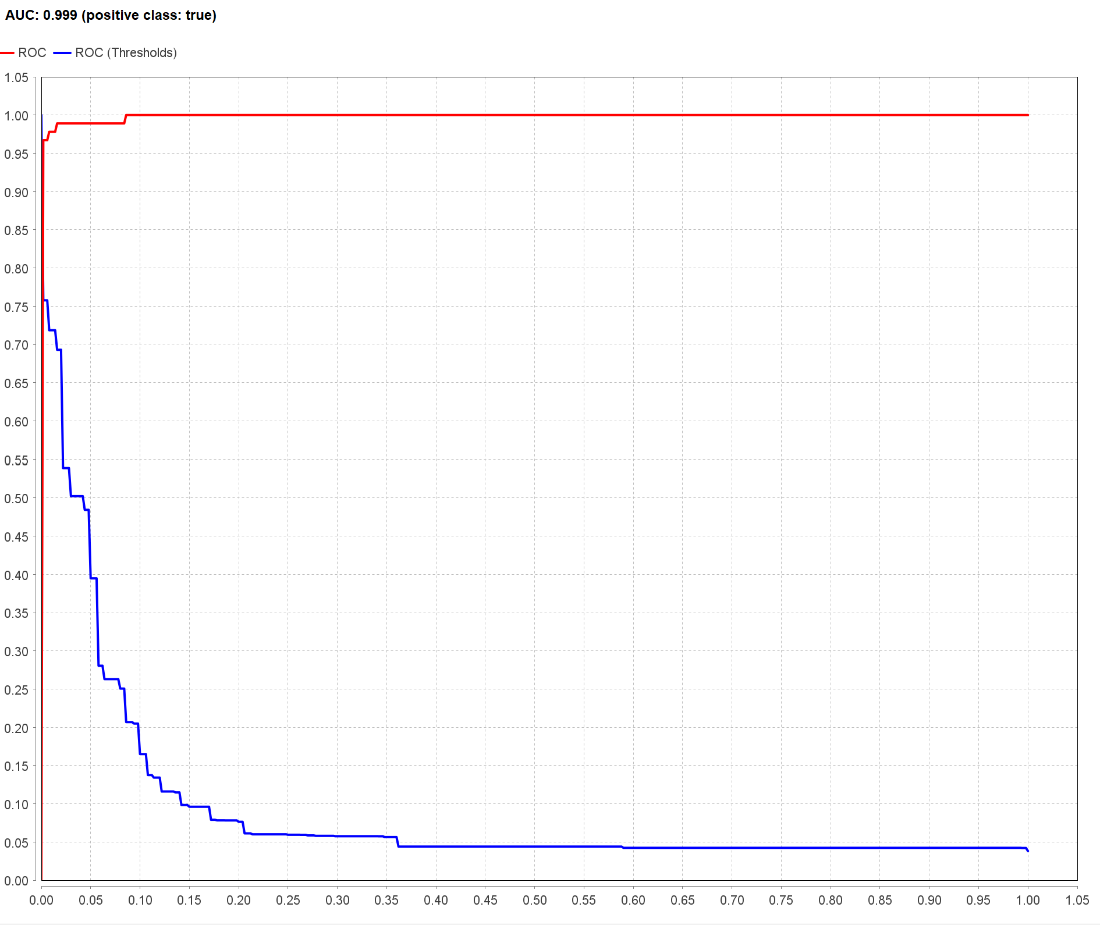
 

## RANDOM FOREST

Our final model selected was the random forest model for the advantages it has over a single decision tree in its ability to classify target variables more accurately. In terms of tuning, the optimal model was achieved through a trial-and-error process in which the criterion and number of trees were manipulated, and performance was evaluated. The final model was created with the criterion set to accuracy and 50 trees. This model predicted the class label more accurately than our k-NN model but less accurately than the logistic regression model. The model returned robust performance metrics in accuracy (98.71%), recall (98.90%) and AUC score (0.999)







# 

# Conclusion

We received a dataset containing 10 variables (one of which was an ID variable not used for modeling) and the binary target variable, which was a flag for either *fraud risk* or *no fraud risk*. This dataset is representative of the challenges that auditors face when allocating their time investigating firms that are likely to be flagged for fraud. There is a need to identify the firms most likely to be flagged as doing so will create a more efficient auditing process.

To begin, we conducted EDA (exploratory data analysis) to uncover relationships and patterns in the data. We found that there are certain district and sector scores that indicate a record will have a higher chance of being flagged for fraud risk. We also discovered that cases with a higher discrepancy between planned spending and actual spending from Reports A and B are far more likely to be cases which were flagged for fraud risk. Additionally, larger monetary values in missed statements indicate a greater tendency toward fraud risk among the cases in the data set.

After identifying these patterns, we developed three different models for classifying the risk flag binary target variable: k-nearest neighbors, logistic regression, and random forest. These models were tuned and evaluated using the accuracy, recall and AUC scores, with the logistic regression model performing best on the validation data set.

We recommend that a classification model be implemented to aid auditors in making more efficient, data-driven decisions when evaluating the companies that could be conducting fraudulent activity. Based on the models we created, we recommend utilizing a logistic binary classification model to achieve the optimal predictions on new case data.

# References

Data Dictionary, <https://guides.library.unt.edu/c.php?g=1307049&p=9616707,>

Excel

Hooda, N., Bawa, S., & Rana, P. S. (2018, April 6). *Full article: Fraudulent firm classification: A case study of an ...* Fraudulent Firm Classification: A Case Study of an External Audit. <https://www.tandfonline.com/doi/full/10.1080/08839514.2018.1451032>