Analysis and Classification of the Audit Dataset

December 25, 2022

Registration number:	2111046
Project:	Analysis and Classification of the Audit Dataset
Link to GitHub:	https://github.com/JJEEEFFFF/CE-888-Assignment-2

Executive summary (max. 250 words)	222
Introduction (max. 600 words)	500
Data (max. 300 words/dataset)	251
Methodology (max. 600 words)	513
Results and Discussion (max. 500 words)	323
Conclusion (max. 500 words)	305
Total word count	2114

Contents

Abstract

Auditors are usually tasked to investigate the accounts of various firms and check whether they are right and do comply with the country's laws or whether they are fraudulent. The aim of this project is to create a machine learning model that can make classification and determine whether a company is fraudulent or not. This is done for the various sectors of the economy which include Irrigation, Public health, Buildings and Roads among others.

The provided data is preprocessed and then used to train three machine learning models, these are: Random Forest classifier, the K-Nearest Neighbor classifier and finally XGBoost. These models are first trained with their default parameters and evaluated by the basis of cross validation. After that, GridSearchCV is carried out to find the best parameters, then the tuned model is evaluated on a different test set.

The accuracies of these models before and after hyper parameter tuning respectively are as follows; Random Forest: 0.9984 and 1.0 after tuning, KNN: 0.9630 and 0.9742 after tuning, XGBoost: 1.0 before and 1.0 after tuning. All the models therefore performed well but relative to each other, the XGBoost model is the best classifier, followed by the Random forest classifier and finally the K-Nearest Neighbor model. The model that can therefore be deployed to production is the XGBoost model due to it's high performance.

1 Introduction

By offering a judgment on whether the financial reports are prepared in all material aspects and in conformity with an acceptable financial reporting framework, the auditor seeks to increase the level of trust held by intended users in the financial statements. High levels of professional knowledge and judgment are necessary for the auditing process. However, it also contains repetitive, time-consuming activities that can be automated by utilizing machine learning as well as artificial intelligence methods.[3]

Process auditing is essential in any firm. The greatest way to assess the effectiveness and general performance of your firm is to conduct timely audits. However, conducting an audit is often an operation that takes time and calls for careful planning.[6]

Auditing is a crucial component of quality management and continuous improvement since it guarantees that your business is operating efficiently and that everything is being managed as it should be. A proper audit also provides a way to determine whether changes are necessary.[2]

The auditing process of a single procedure takes time, let alone hundreds of them spread across numerous facilities. Automation, however, greatly simplifies that procedure. An automated auditing system expedites the audit planning, scheduling, and execution, as well as the audit reporting process. In addition to that, accuracy and consistency are also guaranteed.[4]

The traditional way of auditing has got many drawbacks. This includes the fact that it is reliant on human labor. Human labor is usually prone to getting tired after a while. Human beings are also prone to err since one can unconsciously make a mistake that will lead to a significant impact in the organization. The cost of operation also increases since skilled personnel are required to conduct auditing. This will eventually lead to reduced profit margins in an organization. The other demerit is that the time taken to complete the auditing process will be significantly high compared to when the process is automated. [5]

The traditional auditing method is also based on sampling, in which a random selection of the data is chosen in order to find risks. This is due to how costly and time-consuming it is to analyze all the data. Auditors could come to the incorrect conclusion that there could be more or fewer issues than there truly are if the data they choose is not a sample of the whole dataset under examination. Sampling risk is the term that refers to this issue.[7]

Automation of this process is going to curb the problems that are faced by traditional auditing. This is because it reduces the amount of time spent on time-consuming, repetitive auditing procedures like reporting, data gathering and data extraction. Human ability and knowledge can then be used in additional processes that bring value.[1]

In addition to that, as opposed to sampling, it automatically reviews all accessible data, enabling a more detailed data analysis. Auditors will then have the ability to concentrate on anomalies and records that pose a significant risk as a consequence of having a better grasp of a firm's risk.[1]

2 Data

The data used in the data is the Audit Data dataset from the UCI machine learning repository. This data was collected from The Auditor's office in India from 2015 to 2016 with the purpose of making classifications of firms that were suspicious. On the size of the dataset, it contains 776 instances and 27 features. The data types present are float; 23 features, integer; 3 features and object; 1 feature. For the object type variable, there were three non numerical values that made the entire column to be labeled as Object, so we deleted them since they were very few in number. The correlation between the variables is also shown below:

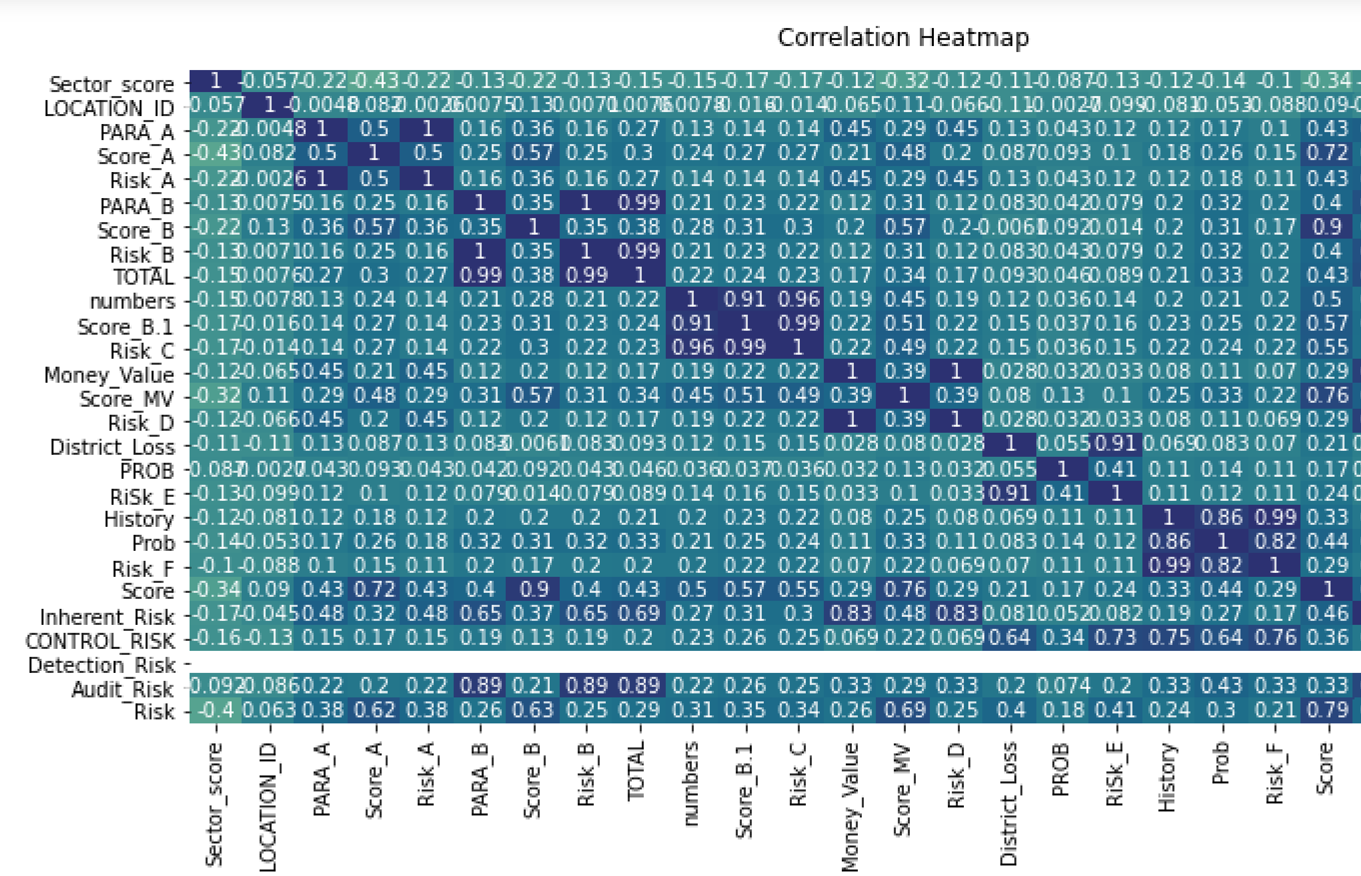

Figure 1: Correlation between variables

The data contains no missing values and therefore no missing value imputation is carried out. Likewise, there are no any other categorical variables hence no dummy variables are created. The data is split to separate dependent and the independent variables and then the data is further split to train and test sets. The ratio used is 80% of the data for training and 20% for testing. The other step is data standardization where the data is scaled to be between 0 and 1. This step is important in ensuring that the model is m=not fitted with values close that are very large as this is erroneous. It also aids in avoiding overfitting and also reduces the model training time.

3 Methodology

The data is first read using the read_csv method of pandas. After this, we check its shape and the data types present. The missing values are also checked. The next step is dealing with the object type variable where we removed the instances of non numerical variables from the column, they were only three.

After this, the data is split into x and y. These are the independent variables and the dependent variable respectively. Further splitting is done in the train-test split using the train_test_split function of sklearn. We choose the size of the training data to be 80% and the remaining 20% to be the test data. It is important to split the data into train and test sets so that part of the data can be used to train the model and the remaining part used to evaluate the model on unseen data. This enables us to know whether the model is good and suitable for deploying, underfitted or overfitted.

On model training, we first used a import a Random Forest model from the ensemble module of sklearn. We import cross validation score from the model_selection module of sklearn so as to evaluate the model. We then use default hyper-parameters to fit it on the standardized data. The model is then evaluated using 10 folds and the average score calculated. We then check for the feature importances after the training is done.

After training with default parameters, we then use Gridsearch CV to tune the model hyper-parameters so that it performs better. The parameters tried out are as follows: "n_estimators": [250, 500, 750, 1000], "max_features" : [4,5,6,7,8,9,10] and "min_samples_split" : [2,3,10]. We then fit Gridsearch on the data and check for the best parameters. A model is then built on these determined parameters, fitted on the training data and then evaluated using 10-fold cross-validation and also on the separate test set. We check for the feature importances after this again.

We next train a K-Nearest neighbor model. We make use of GridsearchCV to find the best hyperparameters of the model. The following are the parameters that are tried out: "n_neighbors": [2,3,4,5], "algorithm" : ['auto', 'ball_tree', 'kd_tree', 'brute'] and "p" : [1,2]. We then fit GridsearchCV on the training data and then check for the best parameters. A model is then created with these best parameters, fitted on the training data and then evaluated using 10-fold cross validation.

The final model trained is an XGBoost model. Again, we make use of GridsearchCV to fine tune the model and find the best parameters. The parameters tried out in this case are 'max_depth' : range(3,10,2), 'gamma' : [0.1, 0.2, 0.3], 'subsample' : [0.8, 0.9], 'colsample_bytree' : [0.8, 0.9] and 'reg_alpha' : [1e-2, 0.1, 1]. We then fit GridsearchCV on the data and check for the best hyperparameters. The model is created with these parameters, fitted on the training data and then evaluated with 10-fold cross-validation.

4 Results

When the random forest classifier trained with default hyperparameters was evaluated with 10-fold cross-validation, it attained an average score of 0.998387. The feature importances for this case is as shown below:

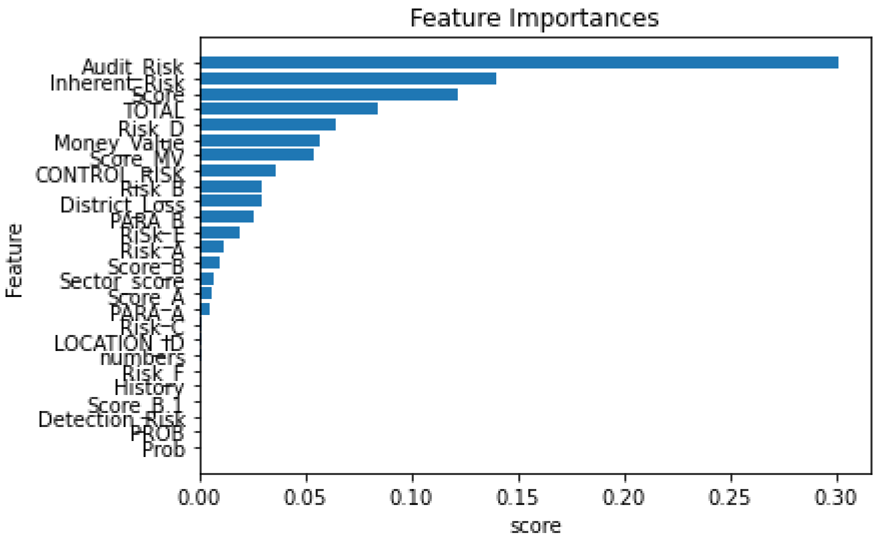

Figure 2: Feature Importances on default hyper-parameters

After parameter tuning, the model attained an average score of 0.9984 on 10-fold cross-validation and 1.0 on the separate test set. The best parameters for the model are 'max_features': 4, 'min_samples_split': 2, and 'n_estimators': 250. The feature importances, in this case, are as shown in the diagram below.

For the KNN classifier, the best model parameters were found to be 'algorithm': 'auto', 'n_neighbors': 5, and 'p': 1. The average score of the model on being evaluated on 10-fold cross-validation is 0.9627. The best parameters for the XGBoost model were as follows 'colsample_bytree': 0.8, 'gamma': 0.1, 'max_depth': 3, 'reg_alpha': 0.01 and 'subsample': 0.8. The average score on being evaluated with 10-fold cross-validation is 1.0

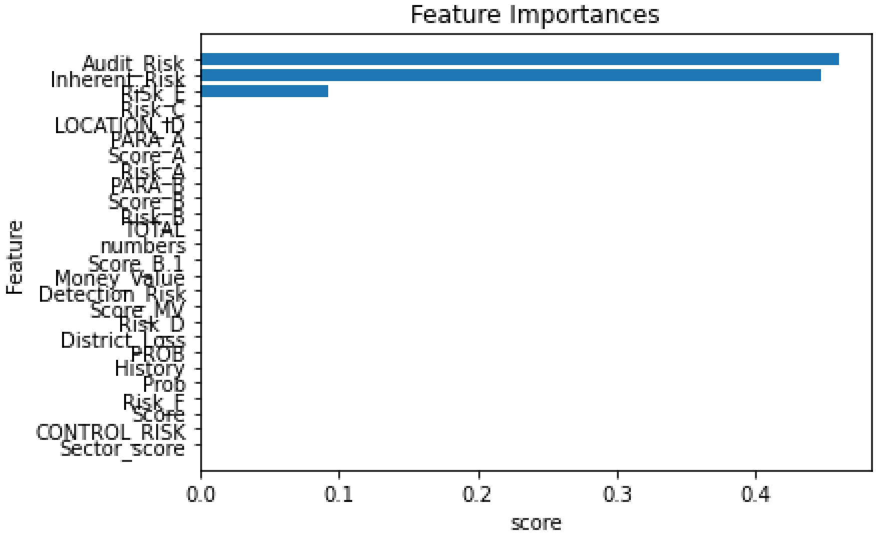

Figure 3: Feature Importances on tuned hyper-parameters

5 Discussion

From the results, it is notable that there are some features with very high correlations with each other (greater than 0.9). These are for instance Risk_A and Para_A, Risk_B and Para_B, Risk_D and Money_Value, Risk_E and District_loss, Score and Score_B. Using only one of these will make not much difference with the results and will also reduce time and effort in collecting the data. The most important features that are dominant in determining Risk are Audit_Risk, Inherent_risk and Risk_E. Special attention must therefore be paid to these features. The best performing model is XGBoost, followed by the Random Forest classifier and finally the K-Nearest Neighbor classifier. The best model to deploy is therefore the XGBoost classifier. To improve the results, we should use more data samples since 776 samples are relatively small. We also need to get rid of the highly correlated features mentioned before and remain with only one of them.

6 Conclusion

In this project, we were able to successfully classify the risk in the Audit dataset. We found the best model to be XGBoost, then the Random Forest and finally the KNN classifier. It was generally a success but it can be improved. This is through the use of a dataset with more samples and also modeling using neural networks.

References

[1] Paul Eric Byrnes, Abdullah Al-Awadhi, Benita Gullvist, Helen Brown-Liburd, Ryan Teeter, J Donald Warren, and Miklos Vasarhelyi. Evolution of auditing: From the traditional approach to the future audit1. In *Continuous auditing*. Emerald Publishing Limited, 2018.

[2] Mark DeFond and Jieying Zhang. A review of archival auditing research. *Journal of accounting and economics*, 58(2-3):275–326, 2014.

[3] Gabe Dickey, S Blanke, and L Seaton. Machine learning in auditing. *The CPA Journal*, pages 16–21, 2019.

[4] Paul A Griffin and Arnold M Wright. Commentaries on big data's importance for accounting and auditing. *Accounting Horizons*, 29(2):377–379, 2015.

[5] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020.

[6] Nopmanee Tepalagul and Ling Lin. Auditor independence and audit quality: A literature review. *Journal of Accounting, Auditing & Finance*, 30(1):101–121, 2015.

[7] J Donald Warren Jr. Embracing the automated audit. *Journal of Accountancy*, 217(4):34, 2014.