# Machine Learning Analysis to Predict Audit Report Cycle Time

*Name removed*

### Air Force Institute of Technology

### *Abstract* — In this project, a machine learning analysis was conducted to determine when a draft report would be issued to management. The project compared the neural network’s performance to the classical modeling approach. The modeling data was obtained from the Air Force Auditing Agency (AFAA) central database and included all reports published between October 2017 and April 2021. The data file included 1,405 completed audits and contained both numerical and categorical data. In this analysis, the mean squared error (MSE) metric was used to measure the performance of the classical and neural network models.

### Overall, the classical model resulted in a 44% chance of accurately predicting when the draft report would be issued with a MSE of 2583 and included 38 input variables. An extensive neural network hyperparameter sweep was conducted, and the best neural network model consisted of 32 neurons, 2 hidden layers, 100 epochs, and a learning rate of 0.1, with a MSE of 2422. Therefore, neither modeling approach achieved the success criteria of predicting when the draft report will be issued to management about 80% of the time. However, this project highlighted significant features in predicting when the draft report will be issued such as the office performing the audit and the report title words of management and military.

*Keywords*  — audit, draft report, machine learning, regression, classical modeling, neural networking, and hyperparameters

### I. Introduction & Background

The Air Force Audit Agency (AFAA) primary performs performance and financial audits. Each performance or financial audit consists of three phases: planning, fieldwork, and reporting. During planning, audit personnel identify primary audit objectives, collect and analyze data, and develop a general understanding of internal control design and effectiveness. Fieldwork allows audit personnel to collect, analyze, interpret, and document information necessary to accomplish the audit objectives and support the audit results. The reporting phase starts when audit personnel prepare a draft report for management comments [1]. The draft report outlines the findings and includes recommendations in order to correct the issues.

Audits help management improve operations by identifying improvements to processes or issues within internal controls. When establishing the annual audit plan, audit personnel work with clients to identity areas of concern they would like addressed. As part of this process, the scope and breadth of the audit are determined to include audit objectives, locations, types of data, availability of auditors, and auditor experience in the topic. An important piece of this process is when the audit results are needed.

It is important to provide timely and relevant results to the clients. Not providing timely audit results could result in outdated recommendations that are no longer relevant or no longer a current issue. Additionally, providing timely and relevant audit results allows clients to implement process improvements to achieve a healthier outcome and provide better mission support to the Air Force. To be of maximum benefit, providing relevant evidence in time to respond to management legitimate needs is the auditors’ goal [2]. Therefore, it is the AFAA’s mission to deliver timely, relevant, and quality audit services [3].

AFAA management requires current, timely, and relevant information in order to make informed decisions. Therefore, predicting when the draft report will be issued to management, clients can request just-in-time audits that are meaningful and have the most impact on their mission. In this project, a machine learning analysis was conducted to determine when a draft report would be issued to management. The project compared the neural network’s performance to the classical modeling approach.

The research question addressed in this work was “using available data, how well can classical machine learning and neural network algorithms predict when the draft report will be issued to management?” Given the input variables, if the classical machine learning and neural network algorithms have the ability to generalize when the draft report will be issued to management about 80% of time that would classify as a success.

Through conducting a lot of research on prior machine learning analyses to predict audit report cycle time, the results concluded that this topic has not yet been studied. However, machine learning has been used in audit to detect fraud or to predict patient scores using regression (see Table 1). In addition, other areas of the audit process are being pursed to implement machine learning and neural network. For example, Deloitte is testing algorithms to read documents such as leases, derivatives contracts, and sales contracts [4]. This would combine the natural language processing with machine learning techniques, which is similar to what this project implemented. Another example is that auditors could use machine learning to detect anomalies and identify emerging risks by looking for patterns of management fraud [5]. A third example is implementing machine learning to improve audit quality by establishing consistency in audit work, such as ensuring a higher volume of work is done accurately without human error [6].

Table 1. Prior machine learning analyses in audit and prediction

|  |  |  |  |
| --- | --- | --- | --- |
| Description of work | Method used | Performance | Ref |
| Detection of fraudulent financial statement reporting | Decision tree | Precision = 0.74  Recall = 0.72 | [7] |
| Classifying self-reported health status in a cohort of patients with heart disease using activity tracker data | Random forest | AUC = 0.79 | [8] |
| Detection of Medicare fraud | Neural network | AUC < 0.70 | [9] |

1. *Data Acquisition*

The AFAA maintains a central database called the Management Information System (MIS). The MIS is used to store all AFAA project management data. The MIS currently maintains information on all the projects performed since 2004. In order to obtain the data used for this project, a query was created in the MIS database. The query included the required data fields along with calculated fields. The calculated fields were used to create the numerical attributes for the dataset needed for this project. As part of the criteria for the dataset, a selection of all reports published between October 2017 and April 2021. That timeframe would reflect AFAA’s most current operations as well as providing enough observations to able to predict the draft report cycle time. Once the query was generated, the data was exported to excel and then converted to a csv file.

1. *Data Understanding*

The dataset used for this project included both numerical and categorical data. The numerical data contained the number of days to complete the planning, fieldwork, and draft report audit phase. The categorical data included the audit office that performed the audit, report titles, and the program type the audit covered. Both the numerical and categorical data required different processes in order to review and understand the data. The review on the numeric data was accomplished with performing statistical analysis. The categorical data required utilizing natural language processes to review the data.

**Numeric data**

Once the csv file was loaded, the data was validated that all the variables were pulled in correctly and each data type was reviewed. The next step was to calculate the descriptive statistics for the numerical attributes. During this review process, the data showed that the planning and fieldwork variables contained minimum values of zero, which means that some of reports did not contain any planning and or fieldwork days and could significantly skew the data. Therefore, these observations were reviewed to see if they should be removed before proceeding with the analysis. In addition, the mean for all three numerical attributes was between 50 and 80. However, the maximum for these attributes were over 370. This reflected that data was significantly skewed (see Table 2).

Table 2. Feature Summary

|  |  |  |
| --- | --- | --- |
| Variable | Observations | Data Distribution |
| Office | 1405 | Categorical |
| Type of Audit | 1405 | Categorical |
| Report Title | 1405 | Categorical |
| Planning Days | 1405 | Skewed Right |
| Fieldwork Days | 1405 | Skewed Right |
| Draft Report Days | 1405 | Skewed Right |

After calculating the descriptive statistics, the next step was to create histograms for each numerical attribute. As shown in Table 2, each numerical attribute is skewed right, which is apparent in the histograms (see Fig. 1). Therefore, these variables were reviewed for possible transformation in order to have a more normal distribution.

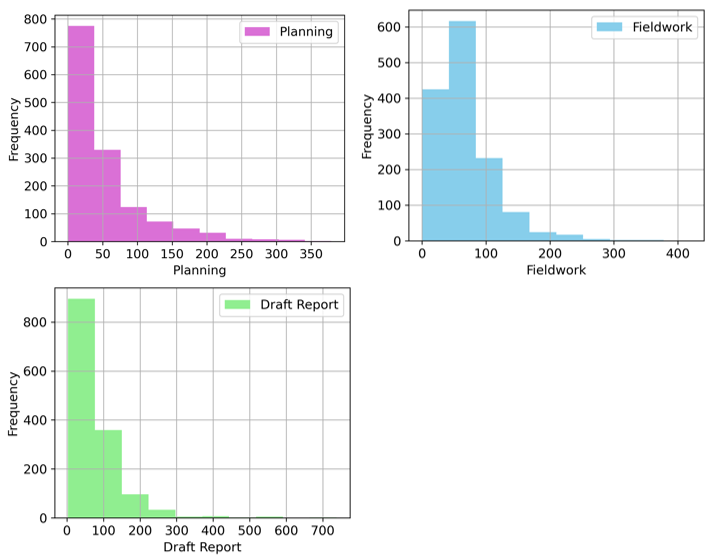


Fig. 1. Histograms of Numeric Variables.

Along with the histograms, scatterplots were created to compare each input numerical attribute to the draft report output variable. When comparing the draft report to the planning and fieldwork variables, the scatterplot revealed no correlations between the input variable and the draft report output variable (see Fig. 2). This means there is no apparent relationship between the planning days and draft report days or between the fieldwork days and draft report days.

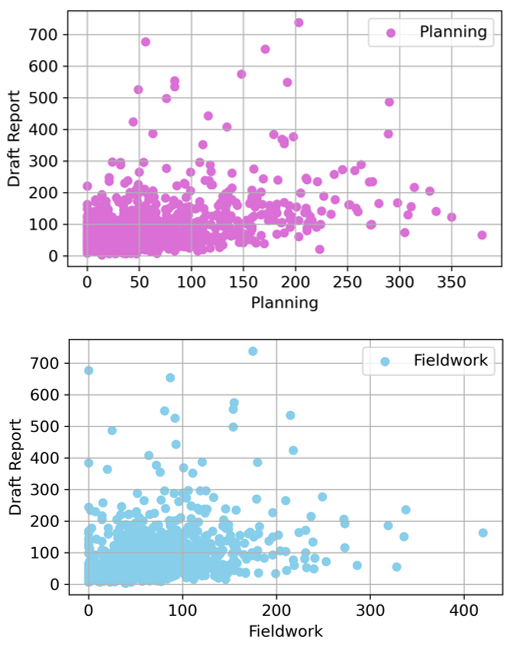


Fig. 2. Planning and Fieldwork vs. Draft Report Scatterplots.

**Natural language data**

The AFAA report titles are important because they describe the type of audit work that was performed or the subject area that was reviewed. For example, if the report title is Security Clearance Inventory, then this was an inventory type audit and could be performed quickly. In addition, if the report title is Nuclear Cyber Security, then this subject area may require more time to review. Each type of audit or subject area can influence the average cycle days until a report is issued to management. In this project, the total number of unique words in the titles was 1,590 and the average word length was 5.9.

In order to include the report titles as part of the modeling process, the titles had to first be standardized. This meant that all non-alphabetical characters had to be remove and the remaining text was in lower case format. Next, tokenization was implemented to create a list of words. Then, all the “stop words” were removed from the list in order to create a list with the most important words. Finally, a graph of the most 12 common words was developed (see Fig. 3). Within the 12 most common words, ten of words all related to the unit or location where the audit was performed. This is because most of the report titles included the unit that was audited and the unit’s location along with the main report title.

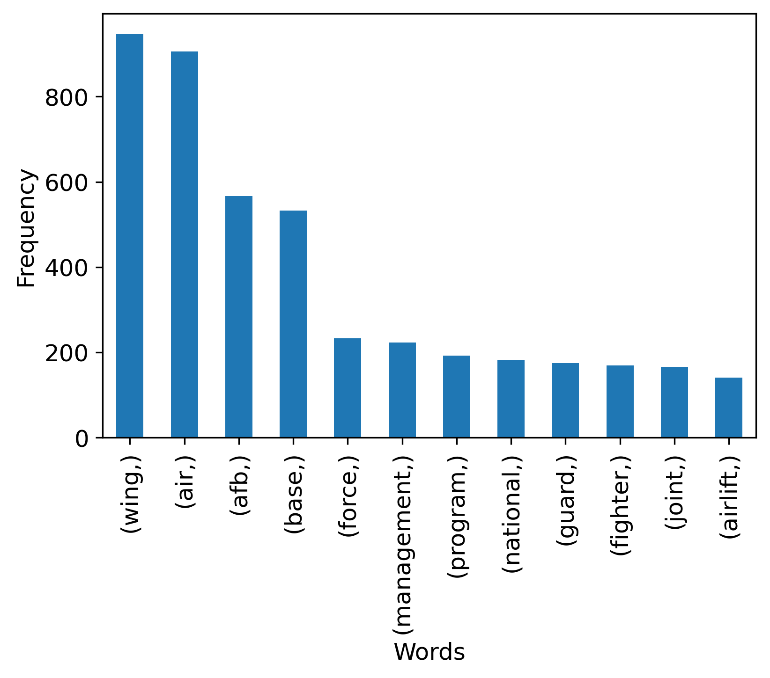


Fig. 3. Report Title Most Common Words.

### II. Method

After completing the data understanding process and reviewing data, it was determined that the data was not quite ready to be used for modeling. Both the numerical and categorical data required work in order to prepare it for modeling.

1. *Data Preparation*

During the review of the histograms, it was determined that the draft report data was significantly skewed right. Since the draft report data was the output variable, it required transformation. The log transformation was used to improve the normal distribution of the data (see Fig. 4).

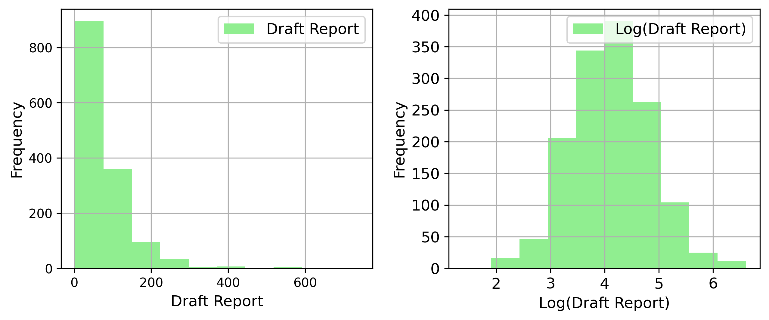


Fig. 4. Draft Report Log Transformation.

In addition to the draft report transformation, the office and report title data required cleaning. The office data included long descriptive names, such as A00900 - Special Programs (FDZ). Therefore, the names were shorten to make the modeling process simpler. For the report title data, all the unit and the unit’s location information was removed from the title. This improved the most common words list by including words that were more meaningful to the audit accomplished, such as management, program, equipment, and access (see Fig. 5).

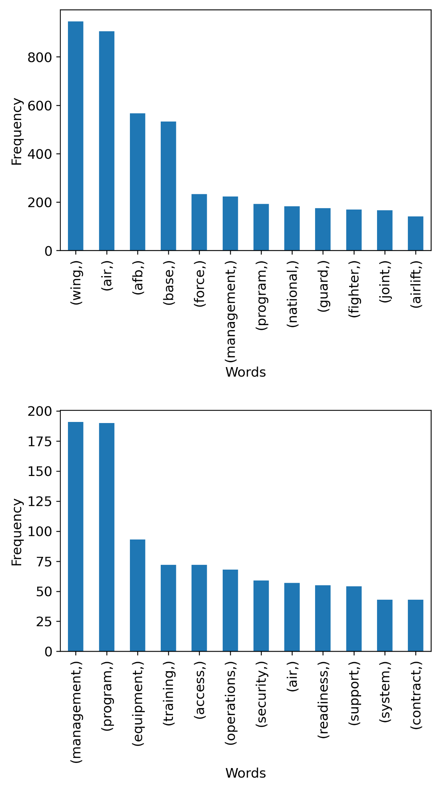


Fig. 5. Report Title Most Common Words Comparison.

Once both the numerical and categorical data was prepared, they were combined to begin the modeling process. The combined dataset consisted of 50 input variables and one output variable. The input variables included 20 of the most common words in the report titles.

1. *Metrics*

In this analysis, the mean squared error (MSE) metric was used to measure the performance of the classical and neural network models. The MSE measures the average squared difference between the estimated values and actual values. It is a significate metric to compare between models. In addition, the R-squared metric was used to measure the performance between the classical models. The R-squared was selected as an important metric because it measures the strength of the relationship between the model and the dependent variable.

1. *Classical Modeling*

Three classical modeling variations were performed. The first variation utilized the recursive feature elimination (RFE) iteration method. This method was performed to easily display the most important features based on the number of variables included in the model. Based on the results, a model with 13 features was used. The second modeling variation was a manual selection and removing input variables based on the p test values. All the report title words with a p test value greater than 0.3 was removed from the model. This model resulted in containing 36 features. The third variation used the draft report log transformation in order to normalize the data.

In all three modeling variations, the fieldwork variable contained a low p test value and was significant in predicting when the draft report will be issued to management. The planning variable was included in two of the models. In both models, the p test value was less than the commonly used 0.05 alpha level, which indicates that it was a significate variable. The categorical variables had a range of p test values. For example, the report title words ranged from 0.02 to 0.97.

In order to prevent overfitting, the data was split into two sets, a training set and a test set. The split resulted in 80% for training and 20% for testing.

1. *Neural Network Modeling*

The neural network modeling included the Adam optimization algorithm, MSE loss function, and the early stopping regularization technique. The Adam optimization algorithm was selected due to the data set containing 50 input variables and requiring multiple of hidden layers. The MSE loss function was included because this project was a regression problem. Further, the early stopping regularization technique was selected in order to prevent overfitting.

Another way to prevent overfitting was to split the data into a training set and a test set. Then, the training set was split for validation. The splits resulted in 80% being training, 20% being testing, and then 20% of the training set was used for validation. The dataset contained 1,405 data points and 50 input variables. Therefore, both the earlier stopping and splitting the dataset helped to prevent overfitting.

The neural network modeling included three variations. The first variation was a preliminary model to use as a baseline without any hyperparameter (HP) testing. The second model was created based on the results from a multiple HP sweep. The sweep tested different combinations of the neurons per layer from 1-32, the number of hidden layers from 1-4, the learning rate from 0.0001-0.1, and the number of epochs from 10-200. The third model was a result of the multiple HP sweep with the early stopping regularization technique included.

### III. Analysis & Results

The results from the classical models and the neural networking models were similar. Both approaches resulted in an average MSE value of about 2500.

1. *Classical Modeling*

All three classical modeling variations resulted in an R-squared value less than 0.5 or 50%, meaning the relationship between the model and the dependent variable is weak. In addition, the MSE for each model is high, meaning the difference between the estimated values and actual values is significant (Table 3).

Table 3. Linear Regression Model Results

|  |  |  |
| --- | --- | --- |
| Model | R-squared | MSE |
| Full | 0.44 | 2601 |
| RFE | 0.36 | 2606 |
| P-value Selection | 0.44 | 2583 |
| Log Transformation | 0.48 | 2731 |

As shown in Table 3, each model resulted in a high MSE, which is apparent in the residual analysis graphs (see Fig. 6). Each model had difficulty predicting when the draft report would be issued to management.

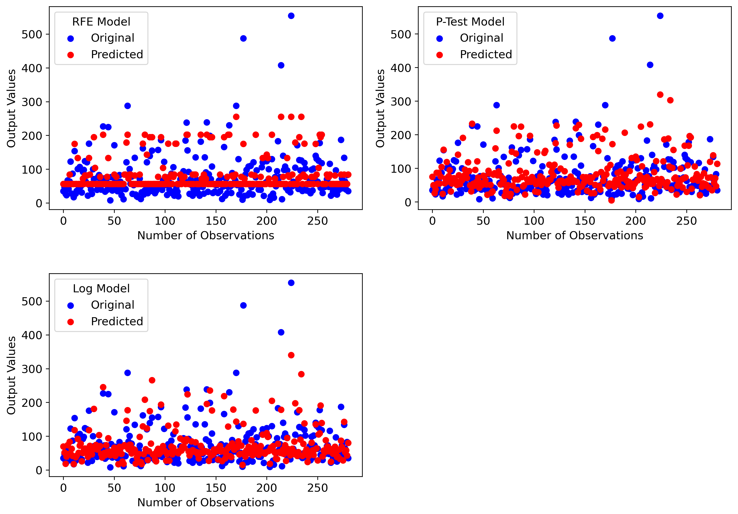


Fig. 6. Original vs. Predicted Model Output Results.

However, even though the R-squared values were low and the MSEs were high, the features included in the models were generally significate based on their p test values (see Table 4). Conclusions about the relationships between the variables and the models can still be made [10]. Therefore, the models are still useful.

Table 4. Linear Regression Model Features

|  |  |  |
| --- | --- | --- |
| Model | Number of Features | Number of Features with <0.05 p test Values |
| Full | 50 | 16 |
| RFE | 13 | 13 |
| P-value Selection | 38 | 19 |
| Log Transformation | 50 | 19 |

The best classical model was the p-value selection model. This is the best model out of the three variations because it has the lowest MSE. In addition, even though it does not result in the best R-squared value, it contains less features than the log transformation model making it less complex.

1. *Neural Network Modeling*

The neural networking models produced a wider range of MSE values than the classical models. The preliminary model resulted in the highest MSE value and the HP model achieved the lowest value (see Table 5). The HP sweep testing revealed that the model achieved the lowest MSE with a learning rate of 0.01 or 0.1. In general, the more neurons, layers, and epochs that was added, the better the MSE. However, the learning rate made the most impact. Based on the results of the first multiple HP sweep, the early stopping included a patience setting equal to 15 epochs because the models seemed to be overfitting quickly.

Table 5. Neural Network Regression Model Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | MSE | Number of Neurons | Number of Hidden Layers | Learning Rate | Epochs |
| Preliminary | 3086 | 4 | 2 | 0.005 | 100 |
| HP Sweep | 1535 | 32 | 3 | 0.01 | 200 |
| Early Stopping | 2422 | 32 | 2 | 0.1 | 100 |

The best neural network model is the HP with early stopping regularization model. This model achieved a lower MSE value than the preliminary model. In addition, even though the HP without early stopping resulted in a lower MSE, it did not account for overfitting. The HP with early stopping regularization model prevents overfitting.

1. *Model* *Evaluation*

The classical model resulted in a 44% chance of accurately predicting when the draft report would be issued with a MSE of 2583. In addition, the neural networking model resulted in a MSE of 2422. Therefore, neither modeling approach achieved the success criteria of predicting when the draft report will be issued to management about 80% of the time. However, this project highlighted significant features in predicting when the draft report will be issued. For example, the office performing the audit was a significant feature. Therefore, there is a relationship between which office performs the audit and when the draft report will be issued. The AFAA can use this information when building the annual plan to determine the number of audits to include for each office to ensure draft reports are issued timely. In addition, the report title words of management and military were significant features and had a relationship with the timely of providing the draft report to management.

Based on the results from the modeling approaches, the selected input variables did not have a strong relationship in order to predict when the draft report would be issued. Therefore, other variables may be needed in order to improve the prediction. In addition, there may be unexplainable variability, such as personnel performance.

The final model selected was the neural networking model with early stopping regularization. This model was selected because it achieved the lowest MSE without overfitting the data.

The HP sweep model achieved the lowest MSE compared to all the other models. However, the model was overfitting the dataset (see Fig. 7).

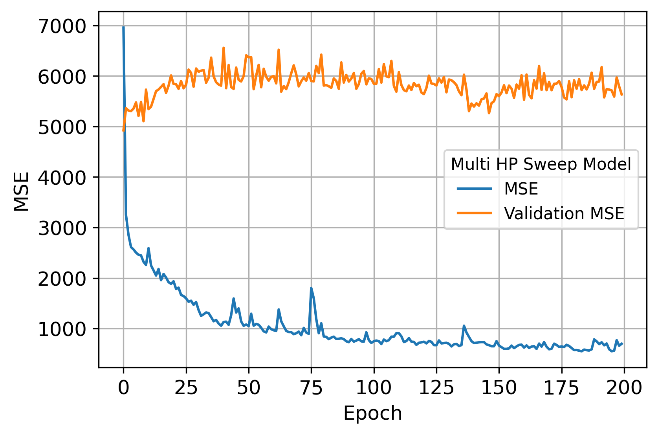


Fig. 7. Neural network overfitting the dataset.

In order to prevent from overfitting the dataset. The early stopping regularization technique was included in the HP sweep. The early stopping regularization technique contained a 15 epoch patience setting to reduce the amount of overfitting. The results are shown in Fig. 8.

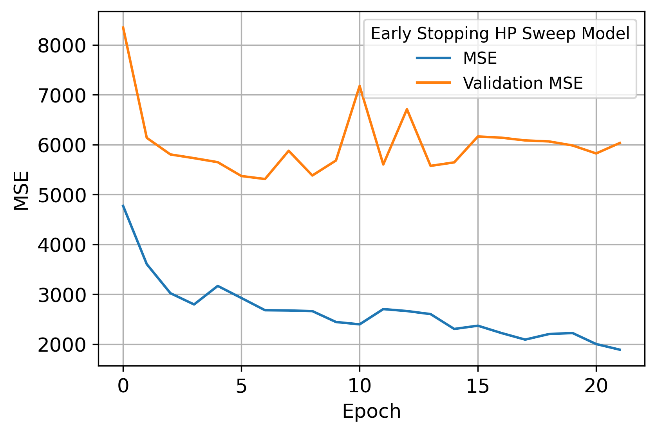


Fig. 8. Overfitting corrected by early stopping.

1. *Model Application (a.k.a. deployment)*

Two different models were used to determine if a model could predict when the draft report would be provided to management from a completed audits dataset.

Overall, the classical model resulted in a 44% chance of accurately predicting when the draft report would be issued with a MSE of 2583.  In addition, the neural networking model resulted in a MSE of 2422.  Therefore, neither modeling approach achieved the success criteria of predicting when the draft report will be issued to management about 80% of the time.

* Qualitative applications:
  + The neural network work model provided the best prediction. However, the classical modeling highlighted significant features in predicting when the draft report will be issued.
  + For example, the project determined there was a relationship between which office performs the audit and when the draft report will be issued.  There was also a relationship between the report title words of management and military and when the draft report will be issued.
  + In addition, a variable that should have been included was the tenure of the auditors or the number of completed audits.  Further, the model should have included a variable for the number of personnel performing the audit. Adjusting the model to include these data points should improve the accuracy.
* Quantitative applications:
  + The current model could assist AFAA management on where to assign Air Force management requested audits during the annual audit plan development.  Also, if an audit is requested by management with a short deadline, using the model to select the office to perform the audit can help ensure a timely results to Air Force management.
  + The mission of AFAA is deliver timely, relevant, and quality audit services. Not providing timely audit results could result in outdated recommendations that are no longer relevant or no longer a current issue. Additionally, providing timely and relevant audit results allows clients to implement process improvements to achieve a healthier outcome and provide better mission support to the Air Force.
  + Using the model to predict when the draft report will be provided to management can be utilized to determine manpower requirements. For example, does AFAA have the necessary manpower to perform the audits in the annual plan?

### IV. Conclusion

Overall, the classical model resulted in a 44% chance of accurately predicting when the draft report would be issued with a MSE of 2583. In addition, the neural networking model resulted in a MSE of 2422. Therefore, neither modeling approach achieved the success criteria of predicting when the draft report will be issued to management about 80% of the time. However, this project highlighted significant features in predicting when the draft report will be issued such as the office performing the audit and the report title words of management and military. Therefore, there is a relationship between which office performs the audit and when the draft report will be issued. The AFAA can use this information when building the annual plan to determine the number of audits to include for each office to ensure draft reports are issued timely. In addition, one of the variables that should have been included is the number of audits the auditor had performed. Senior auditors should complete audits within goals versus new auditors. Further, the model should have included a variable for the number of personnel performing the audit. Consequently, adding these variables may improve the model in predicting when the draft report will be issued.

### References

Note to instructor: I am using the Word bibliography tool.

|  |  |
| --- | --- |
| [1] | Air Force Audit Agency, "Internal Audit Services, AFI65-301," Office of Headquarters Financial Management, Washington, D.C., 2018. |
| [2] | Comptroller General of the United States, "Government Auditing Standards: 2018 Revision Technical Update April 2021," U.S Government Accountability Office, Washington, DC, 2021. |
| [3] | Air Force Audit Agency, "Audit Service Execution, AFAAI65-101," Office of Headquarters Financial Management, Washington, DC, 2018. |
| [4] | G. Dickey and S. Blanke, "Machine Learning in Auditing," June 2019. [Online]. Available: https://www.cpajournal.com/2019/06/19/machine-learning-in-auditing/. [Accessed 9 May 2021]. |
| [5] | IIA, "​The Artificially Intelligent Audit Function," 15 September 2020. [Online]. Available: https://www.riskman.in/2020/09/15/%E2%80%8Bthe-artificially-intelligent-audit-function/. [Accessed 9 May 2021]. |
| [6] | D. Hood, "A new eye on audit quality," 29 March 2021. [Online]. Available: https://www.accountingtoday.com/news/a-new-eye-on-audit-quality. [Accessed 9 May 2021]. |
| [7] | P. Hajek and R. Henriques, "Mining corporate annual reports for intelligent detection of financial staement fraud - A comparative study of machine learning methods," *Knowledge-Based Systems,* vol. 128, pp. 139-152, 2017. |
| [8] | Y. Meng, W. Speier, C. Shufelt, S. Joung, J. Van Eyk, C. N. Merz, M. Lopez, B. Spiegel and C. Arnold, "A Machine Learning Approach to Classifying Self-Reported Health Status in a Cohort of Patients With Heart Disease Using Activity Tracker Data," *IEEE Journal of Biomedical and Health Informatics,* vol. 24, no. 3, pp. 878-884, 2019. |
| [9] | J. Johnson and T. Khoshgoftaar, "Medicare fraud detection using neural networks," *Journal of Big Data,* vol. 6, no. 1, 2019. |
| [10] | E. Dauria, "Looking at R-Squared," 13 May 2019. [Online]. Available: https://medium.com/@erika.dauria/looking-at-r-squared-721252709098. [Accessed 29 May 2021]. |