The main thing is to choose a project that you are passionate about.

The proposal should contain the following parts.

**PROJECT TITLE**

Pitch and Plan for Research project – Text-mining-based machine learning model for financial statement audition

**AIMS**

Financial audits in convention only involve financial statements analysis done by external auditors who must have a sophisticate education background and domain experience. Nevertheless, financial statement audits still require tremendous effort from auditors and leave a loophole for corruption. To alleviate this issue, this project aims at developing a machine learning based model to help auditors detect false accounting. This model will use both financial statements and all other accessible enterprise data, such as sales data, product information and other files generated from the Business Process of the company.

This overall aim can be broken down into 3 objectives: First (i), acquiring sufficient data for training, which may include labeled (fraud or not) financial statements of companies and other information about these companies. Second (ii), build a machine learning model by trying different classifiers including both supervised and unsupervised algorithms. Third (iii), validating and deploying the model on a real-world business case while improving and adjusting it.

**BACKGROUND**

According to PricewaterhouseCoopers’ Forensic Services Practice, the settlement value of lawsuits for accounting misstatement has an increasing trend (Table 1). In fact, in 2007, accounting fraud consists of 14,1% of all economic crime in Australia.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Accounting cases | | | | | |
| Year settled | No. of accounting-related federal lawsuits | No. of financial restatements | Number of settled cases | Average settlement value ($US) |
| 1996-2000 | 106 | 49 | 161 | 18,600,000 |
| 2001 | 123 | 60 | 70 | 23,800,000 |
| 2002 | 167 | 82 | 81 | 17,400,000 |
| 2003 | 120 | 40 | 80 | 27,800,000 |
| 2004 | 132 | 51 | 78 | 34,800,000 |
| 2005 | 87 | 45 | 84 | 90,300,000 |
| 2006 | 64 | 37 | 77 | 74.100,000 |

Table 1. (PricewaterhouseCoopers’ Forensic Services Practice n.d., pp. 41)

Traditional audits rely on manually check financial statements by an auditor, which has many drawbacks. According to Asare, Wright & Zimbelman (2015), there are 6 major difficulties facing auditors. During the audit process, incentives and opportunities for management are hard to be identified with a fixed audit program. In the aspect of auditors, their lack of experience and too much trust in management will lead to the failing of detection.

To address these issues, machine learning techniques are introduced. Sharma & Panigrahi (2013) evaluate 6 classifiers on financial statement fraud detection, which are Multilayer Feed Forward Neural Network (MLFF), Support Vector Machines (SVM), Genetic Programming (GP), Group Method of Data Handling (GMDH), Logistic Regression (LR) and Probabilistic Neural Network (PNN). Their experiments claim that PNN outperforms others, while GP is the second-best classifier. Another study (Hajek & Henriques 2017) found out that ensemble methods are the best in terms of true positive rate, while Bayesian Belief Networks are the best in true negative rate. However, in consideration of fraud statements can adapt to a fixed predictor, the adaptability of the model is required. Zhou & Kapoor (2011) proposed an evolutionary framework that combines the Response Surface Methodology (RSM) with domain knowledge to solve the adaptability issue of the model.

**RESEARCH PROJECT** In this section address the following issues:

• Describe the significance of your project. Why is this an important problem that

needs to be solved?

Stakeholders:

|  |  |
| --- | --- |
| Stakeholder | How the project benefits them |
| Auditor | Ease their job on checking financial statements |
| The management team of companies | Managers can use the prediction model to monitor its business process |
| Stockholders | Stockholders of the company can evaluate the trustworthiness of the company’s financial statement by using this model |
| Banks | Banks can evaluate the credibility of the company through this model |
| Governments | Governments can identify corruptions though this model |
| The Public | The public will benefit by more transparent |

Project Value Statement:

Although previous studies adopted many machine learning techniques into fraud detection for financial statements, they only include supervised learning methods, by which fraudulent labels must be known as prior. This condition is unrealistic in a real business environment. For which, this research proposes an unsupervised learning algorithm which will include a criterion called "originality". This criterion refers to the trustworthiness of a record which may be calculated by its similarity with other records. The proposed model will give the reliability of all records by reference to both the possibility of fraudulence and originality. In addition, as Asare, Wright & Zimbelman (2015) and Sharma & Panigrahi (2013) admitted, text mining techniques will be future research direction in financial statement fraud detection, this research will include text mining techniques, such as Natural Language Processing (NLP) techniques, to reinforce the proposed model. This is because the proposed model will include not only financial statements from companies but also other information, such as news reports, sales reports and product information, that relate to the audited companies to enhance the performance of the model. This report believes the proposed m will combine text-mining techniques and previous financial statements analysis methods to archive high detection rates and high adaptability.

NLP:

As a component artificial intelligence, Natural language processing (NLP) aims at understanding human language by computer. It achieves this purpose by Syntactic Analysis and Semantic Analysis. Syntactic Analysis is one way of extracting meanings from texts. It converts the original sentences into new sentences that align with rules of formal grammar. Another way of extracting meanings from texts is Semantic Analysis, by which meaningfulness of individual words and their combination is checked.

After NLP, the original text files are converted into a standard data file. These files will integrate into the financial statement file from the training set.

Project scope and the Work Breakdown Structure (WBS):

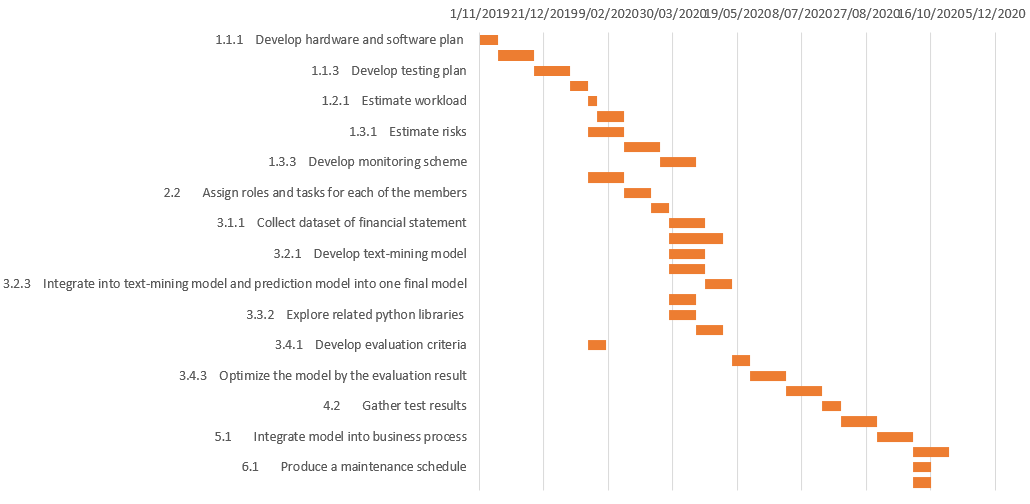
This section discusses all tasks that will be performed in this project and the related work breakdown structure of the proposed project. It is believed that this project will include 6 major scope items, they are plan development, HR management, modeling, testing, deployment and maintenance.

Project:

1. Develop the plan for the project
   1. Develop the technology architecture plan
      1. Develop hardware and software plan
      2. Develop modeling plan
      3. Develop a testing plan
      4. Integrate all plans
      5. The technology architecture plan (Deliverable)
   2. Develop Human Resource management plan
      1. Estimate workload
      2. Develop Positions and Payments for all roles
      3. The Human Resource management plan (Deliverable)
   3. Develop risk management and monitoring plan
      1. Estimate risks
      2. Develop solutions for risks
      3. Develop a monitoring scheme
      4. The risk management and monitoring plan (Deliverable)
   4. The project plan (Deliverable)
2. Human Resource management
   1. Recruit team members
   2. Assign roles and tasks for each of the members
   3. Manage the team
   4. Available team (Deliverable)
3. Modeling
   1. Data collection
      1. Collect dataset of financial statement
      2. Collect other information in the form of text files
      3. Documented data (Deliverable)
   2. Develop the model structure
      1. Develop text-mining model
      2. Develop prediction model
      3. Integrate into text-mining model and prediction model into one final model
      4. Finalized model (Deliverable)
   3. Develop algorithms for the final model
      1. Evaluate state-of-the-art algorithms
      2. Explore related python libraries
      3. Produce new algorithms that suited for the project
      4. Algorithms for models (Deliverable)
   4. Model evaluation
      1. Develop evaluation criteria
      2. Evaluate model under the criteria
      3. Optimize the model by the evaluation result
      4. Optimized model (Deliverable)
4. Test the Prototype
   1. Test the model on a real-business case
   2. Gather test results
   3. Adjust the model according to test results
   4. Usable model (Deliverable)
5. Deploy the Prototype
   1. Integrate model into business process
   2. Train staff for using the prototype
   3. Deployed prototype (Deliverable)
6. Maintain and adapt the Prototype
   1. Produce a maintenance schedule
   2. Produce an adaption plan
   3. Maintain and adapt plan (Deliverable)

Time Estimation

According to the tasks listed above, the time estimation of this project is shown in Fig. xx, which is the Gantt chart of this project. It assumes that this project will start from 1st Nov. 2019 and end by 30th Oct. 2020.



**BUDGET** Give a budget for how much the proposed project will cost the investor. You

should factor in the personnel, equipment and other resources required.

es required.

According to the tasks discussed in the previous section, the estimation of the budget in this project is about 25 million dollars, which can be broken down into 5 components listed in Table 2.

|  |  |
| --- | --- |
| Budget component | $ 1 millions |
| Hardware (on-premise-based and cloud-based) | 100,000 |
| Software (on-premise-based and cloud-based) | 50,000 |
| Project staff cost (Project team payment) | 700,000 |
| Daily expenditure (reimbursement from staff, office replenishment) | 50,000 |
| Contingency | 100,000 |
| Total | 1,000,000 |

**PERSONNEL** Tell the investor what personnel will be required to complete the project and

roughly what their contributions will be.

According to the fact that this project is a research on data mining modeling, the personnel required is estimated in Table 1.

|  |  |  |
| --- | --- | --- |
| Role | Number of people required | Contributions and responsibilities |
| Project manager | 1 | 1. Allocate tasks, monitoring the progress of the project 2. Communicating externally and internally 3. Manage reports and documentations |
| Text-mining experts | 2 | 1. Establish text-mining model 2. Explain text-mining techniques to other team members 3. Integrate the text-mining model into the final model |
| Prediction model experts | 2 | 1. Establish prediction model 2. Explain their model to other team members 3. Integrate the prediction model into the final model |
| Financial audit experts | 2 | 1. Extract business concepts into the modeling process. 2. Provide consultation to technical staff with their domain knowledge 3. Negotiate with buyers |

Because this project does not require procurement and inventory management, technical staff are the major staff in this project. There are 5 roles in the project. First, one project manager is demanded the overall monitoring and control. The manager also has the responsibility to communicate with other external people for acquiring the necessary resources needed in the project. Because the proposed model includes text-mining for enterprise files, two text-mining experts are needed for establishing the text-mining model. Although they will focus on text-mining techniques, they still need to work under the purpose of financial statement audits. Which means they need to communicate with other staff and adjust their model accordingly. The other two technical experts are responsible for establishing the prediction model, which must have the ability to interface with the text-mining model. Finally, two financial audit experts are required for providing support related to domain knowledge. These two staff act as a bridge between the business world and the technical world.

**VIDEO PITCH** A 3–minute pitch that gives an overview of the most important aspects of

the project with the aim of exciting the investors (i.e. me). You should record your pitch and

submit the presentation. Please use online services such as YouTube to host your video

and include the link in your report.

**REFERENCES** List the references for the proposed project. These will be cited probably

from the Background and Research Project sections.

PricewaterhouseCoopers’ Forensic Services Practice, *n.d. Fraud-A guide to its prevention, detection and investigation*, Sydney, viewed 05 October 2019, <https://www.pwc.com.au/consulting/assets/risk-controls/fraud-control-jul08.pdf>.

Asare, S.K., Wright, A. & Zimbelman, M.F. 2015, 'Challenges facing auditors in detecting financial statement fraud: Insights from fraud investigations', *Journal of Forensic and Investigative Accounting*, vol. 7, no. 2,pp. 63-111.

Hajek, P. & Henriques, R. 2017, 'Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods', *Knowledge-Based Systems*, vol. 128, pp. 139-52.

Sharma, A. & Panigrahi, P.K. 2013, 'A review of financial accounting fraud detection based on data mining techniques', *arXiv preprint arXiv:1309.3944*.

Zhou, W. & Kapoor, G. 2011, 'Detecting evolutionary financial statement fraud', *Decision Support Systems*, vol. 50, no. 3,pp. 570-5.