41040 AT2 AI Project

PART: B – REFLECTION REPORT

AUTHORS: Vi Nguyen 13592629

DATE: (3/11/2023)

SEMESTER: (Spring)

YEAR: (2023)

UNIVERSITY OF TECHNOLOGY SYDNEY

LINKS TO SOLUTION NOTEBOOKS:
https://colab.research.google.com/drive/15KLEu-bL0AqUfj2XR02EdkIMQaZmN_KM#scrollTo=NA2NQNJKbCcC

Faculty of
Engineering & IT

Table of Contents

1. Introduction to the project

1.1 Problem and context

A Portuguese bank wants to assess the effectiveness of its direct marketing campaign using machine learning. The target problem is to predict whether the customer of the bank subscribed to a term deposit after the campaign. The data can be retrieved from Kaggle https://www.kaggle.com/datasets/psvishnu/bank-direct-marketing.

1.2 Why customer term deposit is important?

The data was collected between 2009 and 2011 after the 2008 Global Financial Crisis when customers had lost faith in the security of the banking system. The amount of deposits is insufficient for the bank to lend money while maintaining adequate liquidity and complying with the regulations. Therefore, the bank collected the data in an attempt to discover the factors that influence the term deposit of their customers. Maintaining an adequate amount of customer term deposits is critical to the daily operation of the bank.

1.3 Why the project is necessary?

The project was necessary because the analysis considered a large dataset of more than 45,000 rows and 17 attributes. A manual computation and analysis of the relationship between the variables will consume a lot of resources and time. More efficiently, machine learning techniques can discover the pattern to predict whether the customers will sign up for the term deposit based on their attributes. Moreover, similar problems in the future can utilised the model to make predictions with a few modifications.

1.4 Solve the problem

The main approach is to apply the random forest technique to solve the problem. One limitation of this approach is the “Black-Box” feature, which means the program does not produce any algorithm or equation linking the input to the output. The model only provides the final results, which are the correct classification. The technique involved the computation of multiple decision trees at multiple depths, sample split and sample leaf. Each of the trees will try to improve the errors, also known as loss, of the previous tree. The model aims to minimise the loss produced after each ilteration.

A general step to solve the problem included: 1. Import the dataset and all required packages. 2. Explore the dataset and find any potential problem such as missing value, categorical value,… 3. Pre-process and clean the data for further data exploration. 4. Split the dataset into training, testing, and validating. 5. Construct a classifier, train it and make prediction. 6. Evaluate the classifier based on the testing dataset. 7. If the result is not satisfied, perform extra step like parameter tunning, resampling the data in order to improve the performance of the model. 7 Evaluate the classifier based on the validation dataset.

2. Self-reflection **on your personal experience**

2.1 General experience

Overall, I loved the experience of doing this group assignment, even though we encountered many challenges and conflicts. I was assigned as the group leader, so I had to do the task allocation process for my team members. We divided the project into many phases, such as dataset selection, data pre-processing, etc... All the labs we have been learning give us sufficient knowledge to perform these tasks smoothly. We collaborated closely to make a smooth transition between each phase in our model. During the implementation, we had a few conflicts because of different perspectives and backgrounds, but we resolved all the problems and finished the project successfully.

2.1.1 Decisions and actions

The first important decision was the choice of dataset. Our group members came from different backgrounds, so we had different opinions on our desired problem. We then decided to search for a financial data set with the target of predicting customer behaviours such as customer churn or royalty. Considering the current downside of the market, we think it would be more meaningful to predict the bank’s incoming cash flow by classifying whether the customer registers for a term deposit. The data was collected between 2009 and 2011, which was after the period of the Global Financial Crisis (GFC), so we think it can modify the current macroeconomic situation.

Another critical decision was with the AI techniques used. Each team member had different opinions and preferences on the best technique to solve the problem. We then have a meeting with the only purpose is to identify all the advantages and disadvantages of around 10 AI techniques. We then finalise with the random forest technique.

There are a few other important decisions, such as pre-processing the data, hyperparameter tuning and the evaluation metrics to assess the performance of the model.

2.1.2 Barriers encountered

Most problems we encountered required a few investigations except for the imbalance target variable. It significantly reduces the ability of the model to perform the prediction on class “1”, which is the primary purpose of this model. We then have to conduct a lot of further research trying to enhance recall scores. We then finalise with the resampling technique SMOTE and hyperparameter tunning with the target ‘recall’.

2.1.3 Contributions

My main contributions are with the initial code construction, PowerPoint slides and video editing. Regardless, it would be impossible for me to accomplish this project on my own without the effort of my teammates, Bach and Ajwad.

Regarding the code, I had utilised the structure from Lab 4, including the initial process, pre-processing, splitting, training, and evaluating. Additionally, I added the extra two steps of resampling and hyperparameter tuning to achieve the goal of the model. For the slides, I have written the initial script for all the parts, along with the decoration of the slides. As a team, we sat together the next day to adjust and improve my initial scripts. The slides would not be accomplished without the contribution of all team members. For the video, I have recorded the introduction slides and session 4, which is the solution to the problem.

Then, I have put together each contribution into a complete video. Finally, I have submitted Part A, including the code link and the video, while making sure that the format of the submission was flawless.

2.1.4 Group members interactions

Collaboration and continuous communication throughout the project were the key to our success. These interactions allow us to enhance each other’s performance, share our insights and understand different perspectives from different backgrounds. I believe that the diversification of ideas and thinking significantly contributed to the richness of the final project.

2.1.5 Lesson learned: technical and personal.

Personally, I think the most noteworthy lesson was to explore the dataset more carefully next time. I have started with the coding part with an insufficient understanding of the data and its purpose. If I had done more research about each attribute or the documentation of the dataset, then I would know that it was not balanced at the beginning. Hence, I could have further enhanced the overall accuracy and recall score.

From a technical view, this journey of making this AI project has given me a comprehensive view and deepened my understanding of numerous machine learning techniques. Particularly, I have hand-on experience with choosing my problem, comprehending it, and performing all the necessary actions to produce the final solution.

Moreover, I have understood the importance of teamwork in this type of project. I value the diversification in term of thinking and background of each team members. These contributions have increase the quality of the project, widen its horizon and reduce the bias coming soely from my work.

2.1.6 What can be improved?

For future projects, I will try to improve my technical skills by studying more advanced AI techniques and exploring the characteristics of different algorithms. I will also try to employ different models on the dataset to examine the performance if I have the required resources. Moreover, I would love to have more than three members in a team to obtain more valuable views and insights to diversify my project further
Lastly, I think that a more structured approach to project management would have enhanced the quality of our project. For instance, we should have agreed on a fixed day and milestone to do the project. We should also have internal deadlines for each part to enhance efficiency.

2.2 Present three milestone that best demonstrate my learning progression

2.2.1 Milestone I

Understand the fundamental structure of Random Forest in solving the problem and the nature of the problem itself. This includes the process of Explotary Data Analysis (EDA).
The challenges and/or opportunities you met related to reaching Milestone I
I have to understand the strengths and limitations of Random Forest and if this technique is more suitable than others. Moreover, I also have to understand the parameters that impact the predictability of the model. This is a real-world dataset, so we must perform extra research on the year the data was collected. Moreover, we also have to understand why the researchers collected the data in the first place and if there were any limitations or constraints.

I have the opportunity to exacerbate my understanding of the dataset and AI technique, which further improves the performance and trustworthiness of the project.

Resolution strategies that you used to tackle the challenges and/or make use of the opportunities.
To tackle the challenges, I have to refer back to the week 4 resource in Canvas to revise all the information relating to these AI techniques. Additional information was acquired on external websites like sklearn or YouTube videos.

Influence of developing Milestone I in my learning progression
This is one of the most vital steps in the data mining process as it gives me some general knowledge about the dataset. An adequate amount of exploratory data analysis guarantees the quality of the data. Specifically, we can recognise and fix some problems like outliers, data format inconsistency or missing values ultimately enhancing the model's predictability. This process is also a stepping stone into data pre-processing, allowing the model to interpret the dataset quickly.

The appropriate choice of AI technique is compulsory to ensure precise and reliable predictions. For a large dataset like ours, choosing a scalable algorithm like random forest can ultimately reduce the computer resources. Regardless, a complicated technique like a Neural Network may not be the most suitable for our dataset's simplicity. This is where the Random Forest arrives as the optimal solution. The nature of ensemble approach makes Random Forest work very well with this imbalance dataset. The feature importance can also provide insights about the most influential attribute among the 16 attributes, guiding the marketing strategies to deliver better future campaigns.

2.2.2 Milestone II

After exploring the dataset and choosing the optimal algorithm, we focus on implementing the data mining process from start to finish, which include: Data Exploration, Data Pre-processing, Data splitting, Model Construction and Model Evaluation

The challenges and/or opportunities I met related to reaching Milestone I
We could not do data visualisation for the dataset due to technical limitations. Hence, this leads to the difficulty of identifying patterns and relationships between each attribute. The different data types were problematic because Random Forest cannot operate on non-numeric data. Initially, the volume of 45,000 rows was also a big problem because our laptop took intensive training time, especially when it tuned the hyperparameter.

On the positive side, we have the chance to understand the dataset, which can reveal hidden patterns to gain valuable insights for future marketing campaigns. The evaluation stages give us a better understanding of different evaluation metrics than just accuracy, such as F1, recall, etc… We understand that accuracy alone can be misleading, especially for imbalanced datasets.

Strategies that I used to tackle the challenges and/or make use of the opportunities
Instead of visualisation, we implemented the summary statistics of the dataset, counting null, counting the target variable y or checking the data types of the attributes. We employ label encoding to fix the problem of non-numeric data, changing all the instances of string to a number. Regarding the evaluation stages, I have read a lot of documents to understand when and why to use different metrics for model evaluation.

Influence of developing Milestone II in my learning progression
Completing the entire process gives us a better understanding and hands-on experience on real-world datasets. With this experience, I am confident that I can work with any data problem and how to carry out each step to achieve the goals. I have improved my problem-solving skills, able to research to solve my problems. These steps include data cleaning, model constructing and model evaluating. This milestone also further concentrates all the technical skills I have learned during the lab and gives me an opportunity to apply them to solve a real-world problem.

2.2.3 Milestone III

Enhancing the prediction rate of class 1 and improving the recall rate with resampling technique and hyperparameter tuning.

The challenges and/or opportunities I met related to reaching Milestone III
The main challenge was to resolve the unequalled distribution of class 1 (yes) and 0 (no) in the dataset. Label 1 only has around 5000 counts, while label 0 has nearly 40,000. Hence, even though the accuracy is high, it fails to predict our desired target.

Regardless, I have the opportunity to explore numerous techniques to solve this type of problem. I believe that an imbalanced dataset is a common problem in the data mining field. Learning the appropriate technique to solve this type of problem can further improve my technical skills.

Strategies that I used to tackle the challenges and/or make use of the opportunities

First, I have decided to use the resampling technique SMOTE, which oversampled class 1 and undersampled class 0 to make them somewhat balanced. I also investigate the hyperparameters that influence the model's performance, particularly the recall score. I have implemented the grid search to iterate through all combinations for parameters to find the most satisfactory option.

Influence of developing Milestone III in my learning progression

I did not learn these techniques during the formal class and lab. I understand this is a more advanced technique, ultimately expanding my skill set in this field. I recognise the importance and application of the SMOTE technique for future utilisations. With the experience of using hyperparameter tuning, I have acknowledged the importance of parameter selection to achieve a better model performance.

2.3 Reflection of achievements .

Subject learning objectives (SLOs)

2.3.1 Exemplify applications of AI techniques

I have applied the Random Forest algorithm, resampling techniques and parameter tuning to solve the customer term deposit registration problem.

2.3.2 Explain key ideas of common AI techniques:

With my research and fundamental knowledge, I clearly explain the concept of each technique, the advantages and disadvantages of the technique, along with their applications.

2.3.3 Apply common AI techniques to solve simple real-world problems based on existing implementations

I have successfully employed the technique to solve practical, real-world scenarios. The problem required a structured data mining process involving data exploration, data pre-processing, data splitting, model construction and model evaluations. I believe that I can resolve any similar problem outside of the academic environment.

2.3.4. Communicate effectively in both oral and written forms.

- I have communicated my research by composing the video and this report. These two products document my journey and summarise all the essential findings and insights.

CILOs

D1 Technical Proficient

- With my limited skills in data mining and Python, I have improved my technical skills substantially. I have also learned and appreciated a lot of helpful documentation for Python code in the data mining field that I could always refer back to and revise. The code in the notebook, with sufficient comments, no syntax error and high accuracy, are the evidence for my statement.

E.1 Collaborate and communicative

- Collaboration is an essential skill for my future career development. I believe that I have enhanced my communication skills a lot during this project. We must share our ideas and convince the other members to believe it. We have worked together effectively to tackle all the challenges and produce a high-quality project. I believe this experience is a valuable asset and prepared me well before entering the workforce.

2.4 Reflection of your group work experience

2.4.1 Group work experience contribute to your future personal and professional development?

Group work is crucial as each member contributed their skills and ideas to deliver the project effectively. This diversification improved the reliability and performance of our project while removing the bias resulting from just one person's view.

I have also noticed my communication skills develop during this project. Careful evaluation and providing constructive feedback are the most important lessons I have learned. I understand that criticising the idea of the team member will create conflict and affect the final performance of the team.

I have enhanced my skill in translating complex technical terms and evaluation metrics into meaningful business insights. Time management was also a big lesson I have learned. Each member should have their deadlines, and I believe that setting internal deadlines is the key to delivering the project on time.

2.4. If you do differently you had a chance to do this assignment again with the same group?

I love the atmosphere in my current group, where each team member freely contributes their experience and ideas to each other. We have established a no-judgment rule at the beginning, emphasising that all ideas will be considered. While appreciating my current group, I think we can further improve in some aspects.

First, we did the task allocations before understanding the strengths and weaknesses of each team member. We eventually solved the problem by having short meetings to do the task allocations again. Nevertheless, I believe we can reduce the delay in the future if we acknowledge the background of each team member at the beginning.

We have created three different groups for communications, namely Discord, Team and Facebook. I did not understand why we came up with such complications until now, but I will make sure this will not happen again if I have a second chance. We have wasted a lot of time waiting for each other reply on different communication platforms.

2.4.3 What' s your ideal project group for doing this assignment?
My ideal group for this assignment will be people from different backgrounds such as IT, business, marketing, etc… The diversification in the expertise can enhance the performance and the reliability of the model. Moreover, I think communication always creates the most significant impact, so it would be my second requirement. Lastly, I think that a no-judgement environment allow each team member, with suitable expertise and communication skill, to freely express their ideas to contribute toward the final project.

2.4.4 With a chance to do this group assignment in your ideal group, what would you do differently?
My group members already acquired the second and third criteria mentioned in 2.4,3. With a more diverse background, I think that we can make a more significant impact on the marketing campaign of the banking sectors. With feature importance, we can confidently remove the low-score attributes. The removal of irrelevant attributes will enhance the accuracy of the model, reducing computing time and resources.

3 Peer review of a group member

What is the role he/she took in this group project.
We always have regular meetings; therefore, the roles of Bach and Ajwad are not substantially different.

- Contribute toward the dataset selection and exploration.
- Review different techniques and algorithm sto solve the problem.
- Research different evaluation metrics
- Research on what can be improved for the imbalance dataset

Bach:

- Strong in communication and time management
- Able to provide business insights from the code
- Record sessions 3 and 5 for the video
- Analytics skill

Ajwad:

- Record sessions 1 and 2 for the video
- Strong in technical skill
- Coding, providing constructive feedback for my code.

What did he/she do well? What can you learn from him/her?
I can learn a lot from Ajwad's programming background, how to enhance the model performance. I also learned the ability to do clean coding and provide a shorter solution to a similar problem.
For Bach, I am impressed by his outgoing mindset and positive attitude toward our problems. He has connected us, utilising his analysis skills to solve business problems. The business insight produced by Bach sound very professional and suitable for the workplace. I have learned a lot from him.

What can he/she improve?
While Ajwad coding skill is outstanding, I believe that his communication skill can be a limitation for his future career development. We always have to encourage him to speak out his beneficial ideas. Not all parties can understand the technical results like F1 or Recall, so effective communication of the result is critical.
Bach can improve his technical understanding. Even though he does not need to code the entire project, a broader understanding of the algorithm can diversify his skills and make him an influential asset to any team.

What can you do to help him/her improve?
To help Ajwad and Bach's strengths and weaknesses, I have continuously operated weekly meetings where we have each other to improve our weaknesses. Bach and I helped Ajwad communicate the result more effectively while I explained each section of code to Bach with the help of Ajwad.