

Artificial Intelligence Report - Mateusz Jadasz

by Mateusz Jadasz

Submission date: 12-Dec-2024 12:11AM (UTC+0000)

Submission ID: 245663184

File name: 17845_Mateusz_Jadasz_Artificial_Intelligence_Report_-
_Mateusz_Jadasz_621176_1137281860.docx (346.4K)

Word count: 3047

Character count: 16202



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

ML MODEL

by

Mateusz Jadasz

MSc Software Engineering and Artificial Intelligence

Artificial Intelligence

ACCA7026

Seena Joseph

December 2024

TABLE OF CONTENTS

Table of Figures	2
Abstract	3
Introduction.....	3
Problem Statement	3
Methodology	4
Data Preparation	4
Model Development.....	5
Evaluation	5
Experiments.....	6
Results	8
Discussion	9
Conclusion	10
Recommendation	10
References	11

TABLE OF FIGURES

Figure 1: Decision Tree metrics at a 50/50 class weight distribution	6
Figure 2: Random Forest metrics at a 50/50 class weight distribution	7
Figure 3: Decision Tree metrics at a 6/94 class weight distribution	7
Figure 4: Random Forest metrics at a 6/94 class weight distribution	8
Figure 5: Decision Tree Importance Graph	9
Figure 6: Random Forest Importance Graph	9

Demonstration Video Link: <https://youtu.be/q3bGJW3URec>

ABSTRACT

This report examines the application of decision tree and random forest models in predicting term deposit subscriptions based on customer's personal circumstances such as employment status, educational level, etc. The random forest model demonstrated a superior performance, achieving higher recall which was the more desired metric to have a high score in. Challenges such as dataset imbalances were addressed, and evaluation metrics highlighted the effectiveness of ensemble methods in classification tasks. The results of testing highlight the potential of these models to enhance customer acquisition and retention strategies, with recommendations for improving data quality and addressing class imbalances.

INTRODUCTION

In the recent decades, and even more so in the last few years, machine learning algorithms have become essential tools for solving complex problems across various industries. Among those algorithms, the decision tree and random forest models stand out for their interpretability, flexibility, and robustness. This report delves into the process of building and evaluating the decision tree and random forest models for the purpose of predicting term deposit subscriptions using a dataset from a Portuguese banking institution's marketing campaign. Furthermore, this report will highlight and compare their performance, and the insights gained from their application.

By detailing the process of training a model, including data preprocessing, hyperparameter tuning, and evaluation metrics, this report aims to provide a comprehensive overview of how these algorithms can be used to extract meaningful patterns from the data.

The following sections will cover:

- A brief description of the data and its preparation for the models.
- The methodology used for constructing the decision tree and random forest models.
- Key findings, including the models' performance metrics, confusion matrix and feature importance analysis.
- A discussion of the results, limitations, and potential areas for improvement.

PROBLEM STATEMENT

In the competitive banking industry, attracting and retaining customers is critical for sustainable growth. Term deposit subscriptions are one of the key products for financial institutions, offering both customers and banks a secure and consistent investment option. However, identifying potential customers who are most likely to subscribe to a term deposit remains a challenge, even with targeted marketing campaigns.

According to a research paper published in the Journal of Research in Business and Management [1], it would seem that the quality of service is one of the driving factors when it comes to retaining already existing customers, as it determines whether they will continue to use the same service or switch, however in terms of the strategies used for improving customer retention, there doesn't seem to be much of a difference, if any. This suggests that customers are driven more by their own personal needs and whether they are met to a satisfactory level, rather than what a bank or any other financial institution offers and requires. The fact that retention strategies themselves do not bare much weight is further supported by an extract from another

journal, the International Journal of Business and Management [2], where it was found that for the most part, “customer retention strategies were effective to an average extent”. However, although there is no real difference between different retention strategies, and although customers might be more motivated by their personal needs over anything else, customer retention is still extremely important for financial institutions.

With all that in mind, the Portuguese banking institution’s dataset aims to analyse the customer’s personal circumstances instead of looking at the institution’s retention strategies. This is to determine, what exactly are the driving factors for customer retention and which customers should the institution prioritise when it comes to service, time and resources, to maximise the term deposit subscription rate.

METHODOLOGY

To predict the term deposit subscriptions using the decision tree and random forest models, a systematic methodology was employed to ensure accurate and meaningful insights. This process encompassed data preparation, model development, and evaluation. Each step was designed to address the research objectives and account for the complexities of the dataset.

DATA PREPARATION

Before the development of the model could begin, it was essential to not only prepare the data, but to also understand it. By assessing the raw data set, it was clear that it wasn’t originally formatted to just be put into a program and used to build a model straight away, in fact, it was quite messy. This meant that the first thing that needed to be done was a cleanup, so that it was readable. Fortunately, all it took was specifying that each feature of the data set is separated with a semi-colon and that quotation marks were to be removed. Once that was done, the data could be displayed and read in a nice, clean table. Once the data was readable, the analysis could begin.

Firstly, it was important to see exactly what types of data were present in the data set. This was done by summing up all the data types present in the 41188 instances. This was important because the way quantitative data can be analysed is vastly different to how qualitative data is analysed; therefore, it must be separated. Furthermore, although not essential, it was helpful to see how many unique values were present in each feature to get an idea of the variety in the data set.

Secondly, it was crucial to determine whether the data set had any missing values. At a glance, it appeared that some features held the value ‘unknown’, but aside from that it was difficult to tell whether there were any empty features. Therefore, the program was made to look for any features that were ‘unknown’ or even empty, to determine if any data imputation was necessary. After getting a result and finding that no data was missing, only ‘unknown’, it seemed that the best course of action would be to leave it all as it is, as even ‘unknown’ information could still be considered useful; Especially when dealing with data about other people’s personal circumstances which will not always be known should the model ever be put to the test in the real world. Furthermore, replacing ‘unknown’ data with something like a mean, median or mode average could create unrealistic representations of customers and therefore would feed the model incorrect and unrealistic training data, hindering its precision and recall.

Thirdly, the data was ready to be visualised in the form of charts. Qualitative data was visualised in the form of pie charts, whereas quantitative data was visualised in the form histograms. The pie charts allowed for a visual representation of the data distribution and the percentages of each present value. This was useful because it allowed for a better idea of which variables held more sway, and which variables were more prevalent across the data set. The histograms served a similar purpose, however more so for quantitative data. They allowed a

visual representation of the most commonly occurring numbers or ranges of numbers in the dataset, which just like the pie charts, allowed for a clear idea of which features, and which value ranges had more sway when it came to determining if a customer was likely to subscribe to a term deposit.

Fourthly, it was important to begin the search for any correlation between features. The qualitative data was represented in the form of stacked bar charts, which showed the percentage of subscribers with respect to each individual value present in each feature. For example, out of everyone who was contacted and had a mobile phone, just under 20% of those customers were actually subscribed to a term deposit. Similarly, a heatmap was more useful for representing quantitative data because unlike qualitative data, these values are likely to occur in many ranges and variations and will not be the same three or four numbers. Therefore, a heatmap is a more efficient way of representing features with a wide variety of data.

Finally, just before the data could be fed to the model for training and testing, it was important to convert some of the features into dummy features. Dummy variables convert the data held within categorical features into numerical features so that the model can train on numerical data only. Each dummy feature is assigned a number such as educational level = 1, marital status = 2, etc. Fortunately, the implementation of tree-based models also meant that multicollinearity was not a concern. Dummies are very important in this case as during testing an error kept occurring where the program couldn't convert strings to floats, therefore creating dummy variables seemed like the best solution.

MODEL DEVELOPMENT

Once the data was analysed and prepared, the development of the model was by far the simplest step, thanks to the built-in ML libraries in python. Using the Decision Tree and Random Forest classifiers from the 'sklearn' library, training and testing the models was simply a matter specifying which features were the dependant and independent variables, and then splitting the data into a training set and a testing set at an 80/20 ratio. Once that was done, simply parsing the appropriate variables into both functions and outputting the predictions was enough to obtain the information the program was made to provide.

EVALUATION

Finally, after all the analysis, and the training of the model was completed, it was time to evaluate the predictions and results. This was accomplished using a confusion matrix, and for the purpose of the task it seemed much more favourable to have a higher recall over a high precision. This is because mitigating false negatives seemed far more important, seen as a false negative would mean that the company would be losing potential customers by not speaking to the ones that would be much more likely to subscribe to a term deposit. On the contrary, if they happen to speak to someone who is less likely to be interested, they wouldn't lose too much and realistically, could only stand to gain from it in the unlikely event that the customer still subscribes even if they are statistically speaking, less likely to. Therefore, a higher recall is more important than a higher precision.

The recall, precision, and the F1 score values were all displayed alongside the confusion matrix for a better understanding of what the confusion matrix is actually showing. Although the F1 score is present, it's more of a convenience rather than a requirement; As previously stated, recall is far more important than precision and although a high F1 Score by having both of those values quite high would be great, it is definitely not a priority. This is because as previously mentioned, the financial institution can realistically only stand to gain from a lower precision, whereas a low recall could potentially have detrimental outcome from a financial standpoint.

Finally, an analysis of the most important features in each model was shown, in order to determine the driving factors behind customer's decision to subscribe to a term deposit.

EXPERIMENTS

The experiments conducted throughout the development process aimed to evaluate the performance of the decision tree and random forest models in predicting term deposit subscriptions. These experiments were designed to assess the effectiveness of the models under various configurations, determine their predictive accuracy, and identify the most influential features contributing to subscription outcomes.

One big part of improving the model's performance was testing the models with different class weights. In this scenario, class weights play a pivotal role as the data set is extremely unbalanced. In fact, the Y feature of the data set, shows that only 11.265% of the customers are subscribed. This means that the training data for both the customers that would and wouldn't subscribe is vastly imbalanced, therefore creating a problem for the model, as it is very difficult to achieve a high F1 score. As previously discussed in the methodology, the recall is far more important than precision in this scenario, which is fortunate because the fact that the dataset is primarily comprised of customers who wouldn't subscribe to the term deposit means that the recall is exactly the metric that would benefit from this sort of data. However, with the very big difference in the amount of people that would and wouldn't subscribe it was difficult to retain a high recall for the prediction of the customers that would subscribe to the term deposit, in fact if the class weights were distributed equally at a 50/50 ratio, the results were as follows:

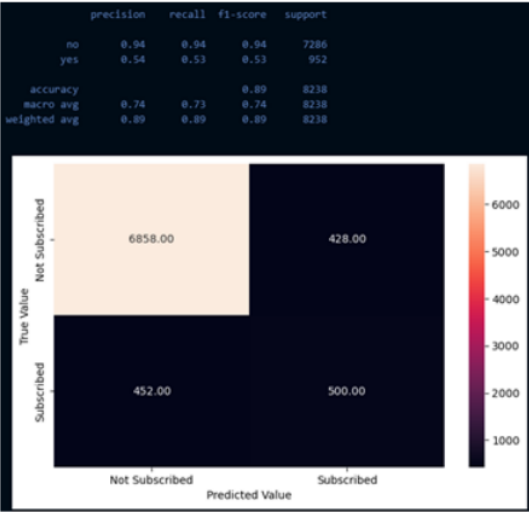


Figure 1: Decision Tree metrics at a 50/50 class weight distribution



Figure 2: Random Forest metrics at a 50/50 class weight distribution

Figures 1 and 2 show that both models had no problem predicting if a customer was likely not to subscribe, however when it came to predicting if they would, was an entirely different matter. With that in mind, applying appropriate class weights was important, to properly reflect the imbalance in data and to try to bring up the recall for the subscribers of the deposit. Multiple class weight configurations were tested, however some of the best results were achieved at a 6/94 ratio split, which really emphasised the importance of the 'yes' value in the Y feature. The results for this split were as follows:



Figure 3: Decision Tree metrics at a 6/94 class weight distribution



Figure 4: Random Forest metrics at a 6/94 class weight distribution

RESULTS

Based on the information shown in figures 1, 2, 3 and 4 in the previous section, one thing was clear, and that is the fact that the Decision Tree model performs better overall when the class weight split is more even whereas the Random Forest works better when the class weight is tailored more towards the data imbalance. This is presumably because the Random Forest is essentially just a collection of Decision Trees working in unison, which means that it has a higher computational complexity. With a higher computational complexity also comes a more specific and accurate set of results therefore a weight distribution tailored to account for the imbalance would make sense to have brought a better set of results in the Random Forest and not the decision tree.

On the other hand, a more balanced set of weights also makes sense to have had brought a better set of results in the Decision tree model as a more balanced set of weights meant that the model could focus on all the data more generally. This performance increase is not very significant in this scenario however, as the changes come mostly from the precision score which as was mentioned previously is not as significant as the recall. In fact, the recall saw absolutely no improvement in the Decision tree model when the class weights were adjusted, it was the Random Forest model that saw a slight improvement in that regard.

Based on the prediction it was important to see which features had the biggest impact on the results. The graphs showing the importance of each feature can be seen below:

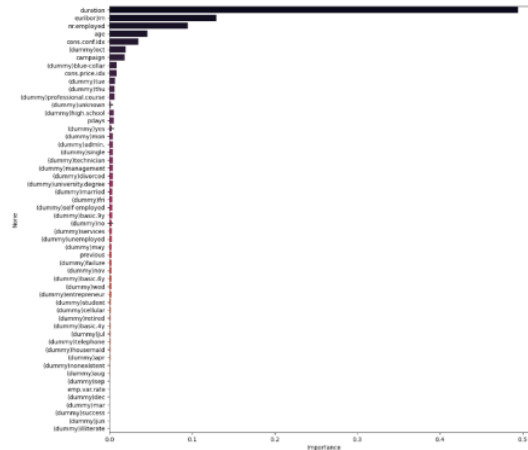


Figure 5: Decision Tree Importance Graph

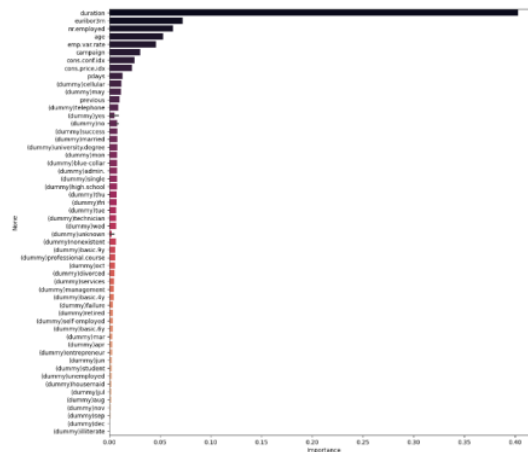


Figure 6: Random Forest Importance Graph

Based on those figures it is evident that the duration feature had the most impact on the overall result followed by the Euribor value, although the Euribor value along with a few others, seemed to have more of an impact on the less complex, Decision Tree. The results demonstrate the effectiveness of the random forest model, outperforming the decision tree model in accuracy and generalizability. These findings align with prior research emphasizing the efficiency and robustness of ensemble learning methods. [3] [4] [5] [6]

CONCLUSION

In conclusion, this report has demonstrated the usefulness and performance of the decision tree and random forest models. The random forest proved to be more robust, using ensemble techniques to achieve a higher accuracy compared to individual decision trees. Key features influencing predictions were identified, providing actionable insights for improving customer targeting strategies. While the findings underscore the effectiveness of ensemble methods in classification tasks, limitations such as the dataset imbalance highlight areas for improvement.

RECOMMENDATION

A recommendation for the future application of classification models like the random forest in scenarios such as this would be to address data imbalances in the dataset, as they can influence model performance and also potentially investing in data collection and quality improvement to expand the dataset, to include more diverse and current customer interactions as they can also enhance model accuracy. Additionally, ensuring the completeness and consistency of data would reduce the need for extensive preprocessing, for example reducing the number of unknown values.

REFERENCES

- [1] R. K. N and Sruthi.V.Devan, "Customer Retention Strategies Used By Banks," *Journal of Research in Business and Management*, vol. 4, no. 8, pp. 91-93, 2016.
- [2] E. K. Mecha, O. Martin and S. M. Ondieki, "Effectiveness of Customer Retention Strategies: A Case of Commercial Banks, Kenya," *International Journal of Business and Management*, vol. 10, no. 10, pp. 270-275, 2015.
- [3] K. Aswini, U. Reddy, A. Nagpal, A. Rana, Praveen and B. S. Z. Abood, "Ensemble Learning Approaches for Big Data Classification Tasks," in *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Gautam Buddha Nagar, India, 2023.
- [4] Y. Ren, L. Zhang and P. Suganthan, "Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 41 - 53, 2016.
- [5] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129-99149, 2022.
- [6] E. Al-Momani, A. Shatnawi, M. Almomani, A. Almomani and M. Alauthman, "A classification model for predicting course outcomes using," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 6, pp. 7090-7102, 2024.

Artificial Intelligence Report - Mateusz Jadasz

GRADEMARK REPORT

FINAL GRADE

81 / 100

GENERAL COMMENTS

Well done!

The report is well written, however it could have been improved by adding the theoretical explanation of both models. Preprocessing and EDA are well presented. The code is well documented. The report lacks the proper explanation about the evaluation metrics and the role and impact of hyperparameter tuning. It could have been improved by adding a table to demonstrate and compare the results of both models and the effect of parameter tuning. Citation and reference are consistent, however the citation could have been improved. For example, instead of indicating, as *"According to a research paper published in the Journal of Research in Business and Management "[1], the better way is "According to the study by Rathi and Sruthi [1] in the reference list, the author's name should not be in initials only.*

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

MATEUSZ JADASZ**81****ENGAGEMENT WITH LIT.SKILL [15]**

*Ability to critically review and integrate relevant and recent literature [6]
Demonstrating understanding of key concepts and theories from literature [6] Providing appropriate references and citations [3]*



Critically reviews and synthesizes recent and relevant literature, demonstrating deep engagement and insightful analysis to support the problem statement. Good understanding of the problem, however, it lacks the review of ML for model predictions. Uses appropriate referencing style consistently, there are minor errors.

13**KNOWLEDGE & UNDERSTANDING [10]**

Depth of understanding of Decision Tree and Random Forest algorithms and their application in predictive modelling (5 Marks) Demonstrating the role of hyperparameters in machine learning models and their impact on model performance (3 Marks) Knowledge of evaluation metrics (e.g., accuracy, precision, recall, F1 Score) for classification models (2 Marks)



Good understanding of the models but lacks depth in theoretical explanation or application.

7

There is no adequate discussion provided for the impact of hyperparameters. However, class weights are adjusted for both models to address the imbalanced data. More parameter tuning is expected to analyse the impact of hyper parameters on model performance.

Mentions metrics but with incomplete descriptions of the evaluation metrics. A detailed explanation of the evaluation metrics is expected.

COGNITIVE & INTELLECTUAL [25]

Critical analysis of the performance of the developed models and interpretation of results (8 Marks) Ability to identify and discuss factors influencing customer subscription decisions based on model insights (10 Marks) Evaluation of strengths and limitations of Decision Tree and Random Forest models in the context of term deposit prediction (7 Marks)



Analyzes model performance but lacks thorough comparison or interpretation. This could have been better by adding a table of comparison and also the effect of tuning.

20

Effectively identifies and explains the factors influencing customer subscription decisions. Discusses strengths and limitations but lacks depth or supporting examples

PRACTICAL APPLICATION [40]

Effective preprocessing of data and explorative data analysis (Eg. handling missing values, encoding categorical variables) (10 Marks) Application of Decision Tree and Random Forest algorithms to the

specific banking dataset (20 Marks) Demonstration of proficiency in model implementation, hyperparameter tuning, and cross-validation (10 Marks)

33



Applies appropriate preprocessing techniques, explaining their impact on data quality and model performance. Implements both models correctly, and applies them effectively to the dataset. The report lacks the proper explanation of tuning and cross-validation. There are evidence of class weight adjustment but it does not connect well with the hyperparameter tuning.

TRANSFERABLE SKILLS [10]

Well organised report (5 Marks) o Clarity and effectiveness in communicating complex technical concepts. o Depth of analysis and understanding demonstrated in exploring predictive modelling techniques and interpreting result. o Alignment of the report content with the initial objectives set for predicting term deposit subscriptions. Demonstrates proficiency in implementing HTML code with well-commented sections and integrating visual presentations effectively (5 Marks)

8



Report is mostly well-organized but lacks clarity in some sections. Code is well-structured, well-commented, and integrates clear visualizations (e.g., plots, confusion matrix) to support analysis