

ZE - 1908686 - assignment

by Zack Evans

Submission date: 11-Dec-2024 09:13PM (UTC+0000)

Submission ID: 245637799

File name: 18253_Zack_Evans_ZE_-_1908686_-_assignment_621176_964545164.docx (428.18K)

Word count: 2896

Character count: 15218



ACCA7026 - Artificial Intelligence

Assignment – ML Model

Student: Zack Evans

Video: <https://www.youtube.com/watch?v=wYgOnTgJ8ml>

Zack Evans
1908686@student.uwtsd.ac.uk

Table of Contents

ACCA7026 - Artificial Intelligence.....	1
Abstract	3
Introduction	3
Methodology.....	4
Data analysis	4
What was conducted	4
Univariate analysis of numerical columns	4
Univariate analysis of categorical variables	5
Bivariate and Multivariate analysis	5
Data preparation	6
Building the models	6
Decision Tree	6
Random Forrest	7
Hyperparameter Tuning	8
Conclusion	9
Recommendation	9
References	10
Figures	11

Abstract

This report focuses on building predictive models to identify potential customers who are likely to subscribe to a term deposit by identifying key factors driving subscriptions and profiling target customer groups. The report applies decision tree and random forest machine learning models to develop a system for the task at hand. Data exploration, preprocessing, and exploratory data analysis is needed to assure the quality of the data is high enough to identify the patterns that drive term deposit subscriptions. The findings compare the strengths and weaknesses of both models and offer recommendations to increase customer subscriptions.

Introduction

Understanding customer behaviour is critical when it comes to creating success within banking and financial organisations. In an increasingly competitive and resources-constrained environment, these organisations must adopt a more data driven approach to optimise marketing strategies and improve campaign outcomes. The ability to identify potential customers for specific products (like term deposits), enables organisations to maximise return on customers.

To do this machine learning models such as decision tree and random forest have emerged as powerful tools when it comes to prediction tasks. Depending on how these models are trained, they can be highly accurate tools which can give the needed insight into customer data.

This report aims to develop a couple of predictive models using both decision tree and random forest to analyse a dataset derived from a Portuguese banking institutions marketing campaign. The primary objective is to predict whether a customer will subscribe to term deposit based on the variables within the dataset. In doing so, the report seeks to uncover the key factors that drive the customers to subscribing to the term deposits so a campaign can be target to customers that would respond positively to the marketing efforts.

To achieve the reports goals, the report undertakes:

- Data exploration and preprocessing
- Exploratory data analysis
- Development and evaluation of decision tree and random forest
- Feature importance analysis
- Comparison of both models

The findings will provide a valuable insight which will enhance marketing strategies and improve the overall campaigns efficiency when it comes to getting customers to subscribe to the term deposit.

Zack Evans
1908686@student.uwtsd.ac.uk

Methodology

Data analysis

Data analysis is the systematic analysis of a body of data. It allows a rounder way of to effectively interpret and conduct hypothesis on data. As I.J. Good explains in his exploration of exploratory data analysis [1], this process balances between descriptive statistics, which is focused on clear presentation and summarisation of data. While also using more graphical representation to hold more in depth understanding on the data to allow hypotheses to be generated on observed patterns.

What was conducted

Within this section we will be looking over the dataset and preprocessing it. To do this a library called panda will be used to load the data and analyse it. Once loaded, it was observed that there were 21 fields, with none of them being unique identifiers.

It was found that two fields called "duration" and "euribor3m" both contained a fair few unique value. But due to the data type being either a float or integer value they were kept due to them not being unique identifiers and possibly valuable when making the prediction models.

Univariate analysis of numerical columns

To analyse the data, the variables were split between numerical and categorical columns. This would make analysing the data easier to do. The main notable observations found during panda describe analysis were:

- Average age is around 40 years. It has a high range, from 17 years to 98, indicating good age diversity in the contacts.
- On average, the number of contacts performed during the campaign was 2 – 3
- The pdays is quite volatile as the default value is 999
- The average contact before the campaign is 0.2 meaning that the bank rarely contacts customers outside of campaigns

Next is to explore the variables in more depth by observing their distributions, please look at [F1] for more information. What was observed from this was:

- The age distribution is close to a normal distribution with most contacts being between the ages of 25 and 50
- Most of the time, the number of times that a customer was contacted during a campaign was in the single digits
- No good analysis can come from pdays, as most values use the default 999 value
- On average, a contact was contacted 0 - 2 times before the campaign

- The employment variation rate has a high frequency of 1 and -2
- Consumer price index has an average of 93 - 94, with a peak just before 94
- Consumer confidence index has several peaks with the average laying around -40
- Euribor 3 month rate has a gap between 2 - 3.5, meaning most values lay either before or after the gap

Univariate analysis of categorical variables

To analyse the categorical variables, each value was analysed to show the proportion of each value within the variable. Observations found from this were:

- The success rate of the campaign was 11%
- Most people are married with it being 60%
- Education wise, most people either have a university degree, or have only finished high school
- Most contacts have more IT focused jobs, with 40% having either admin or technician jobs
- Majority, 99%+ either no or unknown for default
- 52% of contacts have housing loans
- 82% of contacts have loans
- Majority of contacts, use cellular over a telephone
- The campaign was mostly conducted in the first quarter of the year
- There was nothing notable with the day of the week, as it had an even distribution

Bivariate and Multivariate analysis

Now that the numerical and categorical variables have been analysed, we can now check how the subscribe rate is related with other categorical variables. From the observations viewed:

- Even though there is no distinguishment between the days of the week, the later months in the year (sep, oct, dec) have a higher chance of having contacts subscribe to the term deposit with one outlier being March.
- There is a higher success of having contacts subscribe through cellular compared to telephone.
- Clients that have not defaulted are more likely to subscribe with about 15%~
- People who are retired, and students are more likely to subscribe with around 25%~ having subscribed on average compared to the average of 10% for other jobs
- Marital status, education, general loans, and housing loans don't affect the likelihood of people subscribing
- poutcome does outline that, if a previous campaign was successful with the customer, it will normally work again. With around >60%

The next bit of analysis that was conducted was to check the relationship between subscriptions and numerical variables, it was found that:

- Customers that have a lower employment rate are more likely to subscribe
- The same goes for euribor3m

Data preparation

“Machine learning algorithms automatically extract knowledge from machine-readable information. Unfortunately, their success is usually dependant on the quality of the data that they operate on. If the data is inadequate, or contains extraneous and irrelevant information, machine learning algorithms may produce less accurate and less understandable results or may fail to discover anything of use at all.” [2].

This is why data pre-processing is important in terms of machine learning. It is a necessary step to resolve several types of problems such as missing values, and noisy values such as categorical fields. This normally leads to hight quality results, and less resources being used up when data mining.

Implementation

To get the data ready for the prediction models, the data need to be pre-processed so that the models know what they are looking at. Numerical variables are already ready to go as it's easy to train a data model from the data, but the categorical variables need to be processed so dummy variables replace the categorical values.

To do this we need to look at the preexisting data and set dummy integer values to each of them. Within [F3], it is shown that the categories, job, marital, education, contact, poitcome, default, housing, loan, y, month, and day_of_week have been given dummy values. More simple categorical variables like default are Boolean based, so setting it to either 0 or 1 is easy. But for more complex variables like job or education we allow the system to automatically set its dummy values.

Next, we need to drop the variable that shows the customers have subscribed to term deposits. The reason behind this is to create two sets of data. One for training, and the other for testing. The split we will be using is 70% to 30%. 70% for training while the 30 % is used for testing. To reduce any negative impacts when distributing the data between both sets, stratified sampling will be used to ensure that the relative class frequencies are approximately preserved so that the training and testing datasets are reliable.

Building the models

Decision Tree

A decision tree is a supervised learning algorithm which can be used for either classification or regression tasks. “Decision trees are flowchart graphs that show us the possible classification of data, and it can be researched by mathematical methods.” [3]

Zack Evans
1908686@student.uwtsd.ac.uk

Our task is to predict whether a customer would subscribe to term deposit. So, to do this Decision Tree and Random forest were picked due to their features and relevance to the project.

Implementation

In this practical, DecisionTreeClassifier was used from sklearn. We will be supplying two parameters to create the Decision Tree instance. Class weights, and random state. Class weights are a hyperparameter for the Decision Tree, and in this case, we will be setting it to 17:83. This will lead to class B being the dominant class and the decision tree will lean more towards class B. The other parameter random state, we will set it to 1 so that the same result is given even if we run the model multiple times. This is done so that comparisons can be done easier.

After fitting and predicting our training data. We get our results [F4]. As we can see we get a precision, recall and f1 score of 1. Meaning that with the training data and comparing it to the training set we get 100% accuracy.

So, what do these results mean?

- The accuracy means the overall correctness
- Precision evaluates the positive prediction quality
- Recall accesses sensitivity of which positive instances were classified
- F1 score, allows us to identify the balance between precision and recall

In simple terms, as long as the values are higher, it means that the model is performing well.

To see how successful the overall training was, we now check our model against the testing set of data. From results [F5] we can see that our accuracy goes down to 89% with the predicted sub giving many false positives. For the positive prediction we get a precision of 52%, 49% recall and f1 score of 50%. The results show us that when comparing the test and training data, we have a 50% hit to our positive prediction, while the negative prediction is fairly accurate. This could mean a couple of things, either there's not enough positive instances within the dataset to allow the model to train more efficiently, or there is something wrong with the data preparation.

Feature importance

To see what mainly impacts the performance of the model, we can create a feature importance chart which will show us which variables can be used to predict subscriptions.

It is shown that durations, nr.employment, euribor3m, age are the main variables that help predict subscriptions.

Random Forrest

A Random Forest is an ensemble learning method, meaning: "The idea behind the random forest ensemble learning algorithm is to create a forest of decision trees with

Zack Evans
1908686@student.uwtsd.ac.uk

the goal of reducing the variance in the results" [4]. From the group of Decision Trees, a group decision is made to predict the data given.

This in theory should improve the performance the accuracy of our model [5]. But the downside to this is the longer compute time needed to execute it.

Implementation

To implement the random forest, the RandomForestClassifier from sklearn will be used. Just like the decision tree, we use weights and random state as our parameters which will be the exact same.

After fitting and predicting our training data, our results [F6] show a 100% accuracy when doing metric score against the testing data. To get the real results we shall look further into the testing data. From our results [F7], we can see we get an overall accuracy of 92% which is more accurate compared to the Decision Tree. Both models perform poorly with both giving a lot of false positives, but overall Random Forest beats the Decision Tree in precision, recall, f1 score and accuracy, even in giving positive predictions.

One way we can solve the issue with false positives being a recurrence is to preform hyperparameter tuning. The use of hyperparameter tuning can lead to increased performance of the model reducing both overfitting and underfitting.

Feature importance

When conducting the feature performance of the Random Forest, a few similarities were found. Like the Decision Tree, age, euribor3m, nr.employed were found to be important factors when predicting. But the campaign variable was a new addition which appeared.

Hyperparameter Tuning

Hyperparameter tuning can greatly affect the predictive performance of a machine learning algorithm [6]. When conducting hyperparameter tuning on ML models, using a suitable configuration usually done by trial and error. It can be very resource consuming, but the result can lead to better performing models.

This is normally carried out based on a performance measure. In this instance, we want to increase the precision, recall, and f1 score of the positive predictions within the models. To prove this concept, we are going to apply hyperparameter tuning against our Random Forest. We will need to supply the same parameters as before, but with a few new additions. First are the n_estimators, which allows us to control the amount of Decision Trees we want to be developed for the random forest. Next, is the minimum_samples_leaf which allows us to specify the minimum number of samples required to be present in a leaf node. Lastly, are the max_features allows us to control the size of the randomly generated subsets of features.

With the understanding behind the parameters, we can run the hyperparameter tuning and try to improve our model. After running, our results we did see an

Zack Evans
1908686@student.uwtsd.ac.uk

increase in performance with positive detection, but a slight decrease in overall accuracy. But this does mean that we are moving in the right direction when it comes to improving the overall accuracy of the model.

Need to finish

Conclusion

In conclusion, the project conducted did achieve its goals of identifying what type of customers are more likely to subscribe to the term deposit. From data analysis we found out that occupation's such as retired and students were more likely to be subscribed to term deposits. To further ingrain this conclusion, the models identified that the employment fields were indeed very important when it came to predicting.

With this, the organisation can focus more on specific groups that are more likely to subscribe to term deposits, which can result in a decrease of overall resources and an increase of term deposits.

Recommendation

In the future, I think it would be great to look at other models like Naïve Bayes, Logistic regression, and Linear regression to compare each model to the dataset instead of Decision Tree and Random Forest which is fundamentally the same.

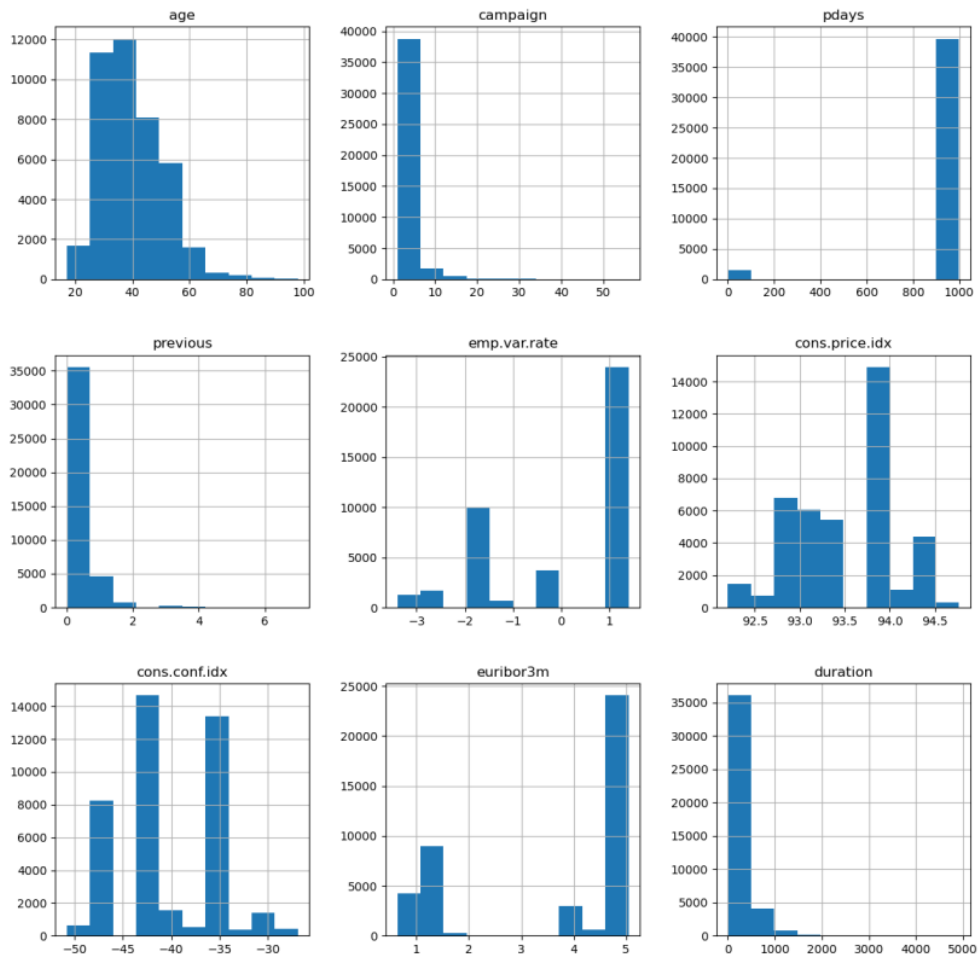
As well as, look further into the performance of each model with different data features being supplied, and how numerical and categorical features affect each one.

References

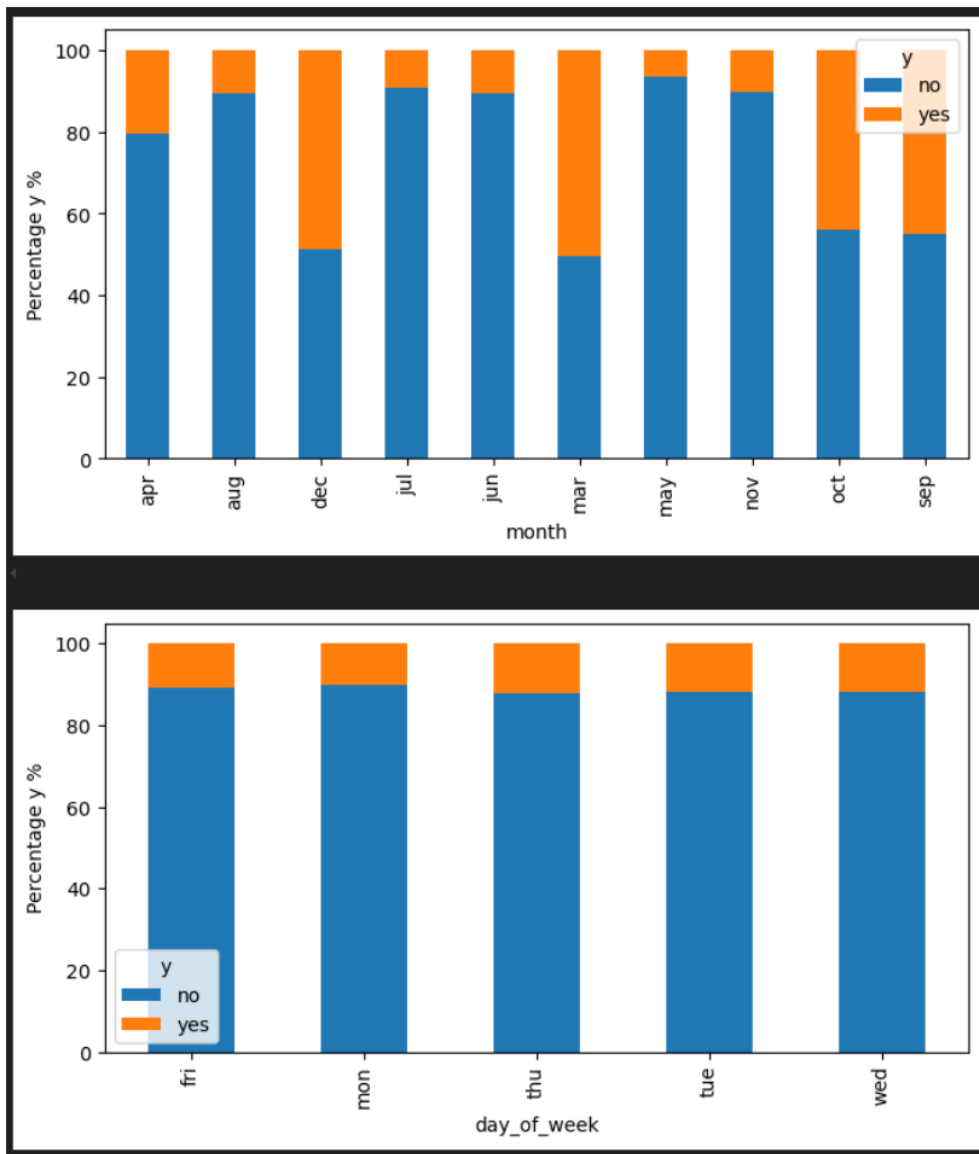
- [1] Good, I. J. (1983). The Philosophy of Exploratory Data Analysis. *Philosophy of Science.*, 50(2), 283–295. <https://doi.org/10.1086/289110>
- [2] Kotsiantis, Sotiris & Kanellopoulos, Dimitris & Pintelas, P.. (2006). Data Preprocessing for Supervised Learning. *International Journal of Computer Science.* 1. 111-117.
https://www.researchgate.net/publication/228084519_Data_Preprocessing_for_Supervised_Learning
- [3] Y. Lu, T. Ye, and J. Zheng, "Decision Tree Algorithm in Machine Learning," *IEEE Xplore*, Aug. 01, 2022. <https://ieeexplore.ieee.org/document/9918857> (accessed Apr. 27, 2023).
- [4] Z. E. Ashari and S. L. Broschat, "t-Tree and t-Forest: Decision Tree and Random Forest Algorithms Including the Relevance Factor with Applications in Bioinformatics," *IEEE Xplore*, Nov. 01, 2019.
<https://ieeexplore.ieee.org/document/8983065> (accessed Dec. 02, 2021).
- [5] P. T. R, "A Comparative Study on Decision Tree and Random Forest Using R Tool," *IJARCCCE*, vol. 4, no. 1, pp. 196–199, Jan. 2015, doi: <https://doi.org/10.17148/ijarcce.2015.4142>.
- [6] R. G. Mantovani, T. Horváth, R. Cerri, J. Vanschoren and A. C. P. L. F. d. Carvalho, "Hyper-Parameter Tuning of a Decision Tree Induction Algorithm," 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), Recife, Brazil, 2016, pp. 37-42, doi: 10.1109/BRACIS.2016.018.

Figures

[F1] Numerical column distribution



[F2] Month based analysis vs day based analysis



[F3] Setting dummy variables

```

#creating list of dummy columns
to_get_dummies_for = [
    'job',
    'marital',
    'education',
    'contact'
]

# Creating dummy variables
df = pd.get_dummies(data = df, columns=to_get_dummies_for, drop_first=True)

# Mapping overtime and attrition
yesNo_MAP = {'yes': 1, 'no': 0, 'unknown': -1}
poutcome = {
    'failure': -1,
    'nonexistent': 0,
    'success': 1
}
months = {
    'mar': 3,
    'apr': 4,
    'may': 5,
    'jun': 6,
    'jul': 7,
    'aug': 8,
    'sep': 9,
    'oct': 10,
    'nov': 11,
    'dec': 12
}
days = {
    'mon': 1,
    'tue': 2,
    'wed': 3,
    'thu': 4,
    'fri': 5
}

df['poutcome'] = df['poutcome'].map(poutcome)
df['default'] = df['default'].map(yesNo_MAP)
df['housing'] = df['housing'].map(yesNo_MAP)
df['loan'] = df['loan'].map(yesNo_MAP)
df['y'] = df['y'].map(yesNo_MAP)
df['month'] = df['month'].map(months)
df['day_of_week'] = df['day_of_week'].map(days)

```

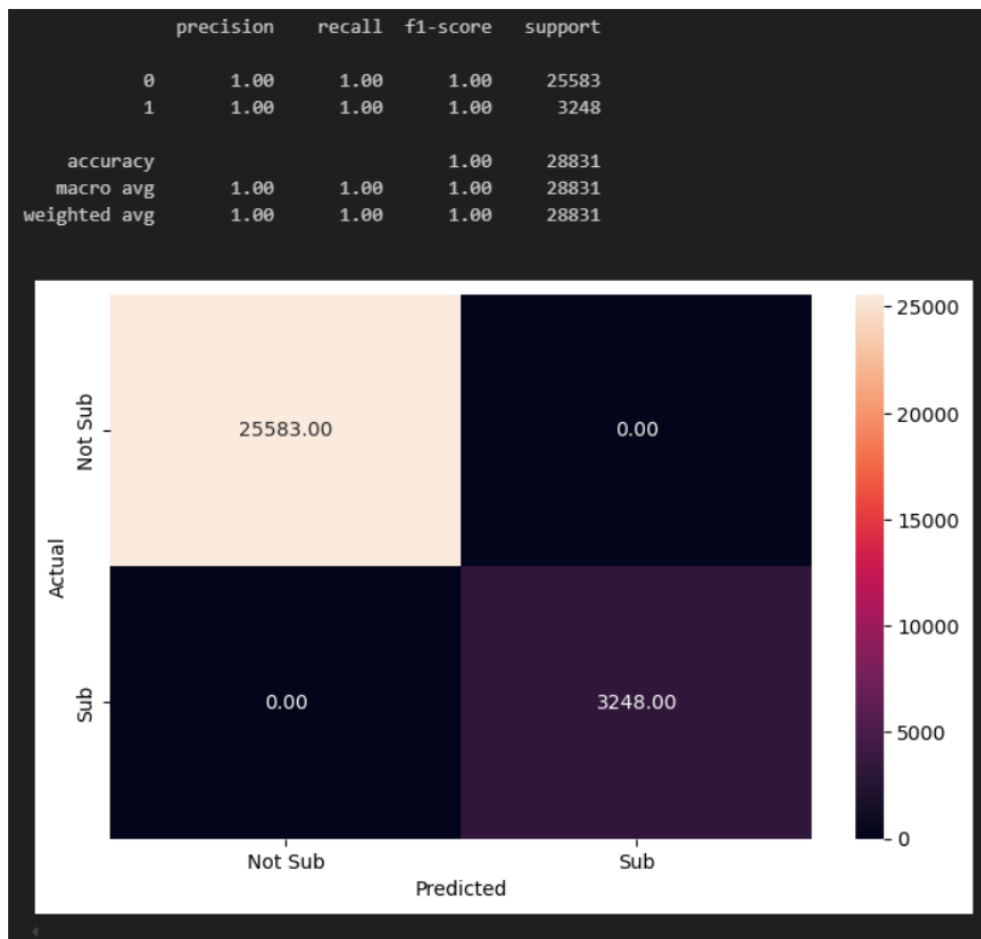
[F4] Decision Tree performance with training set



[F5] Decision tree performance with test set



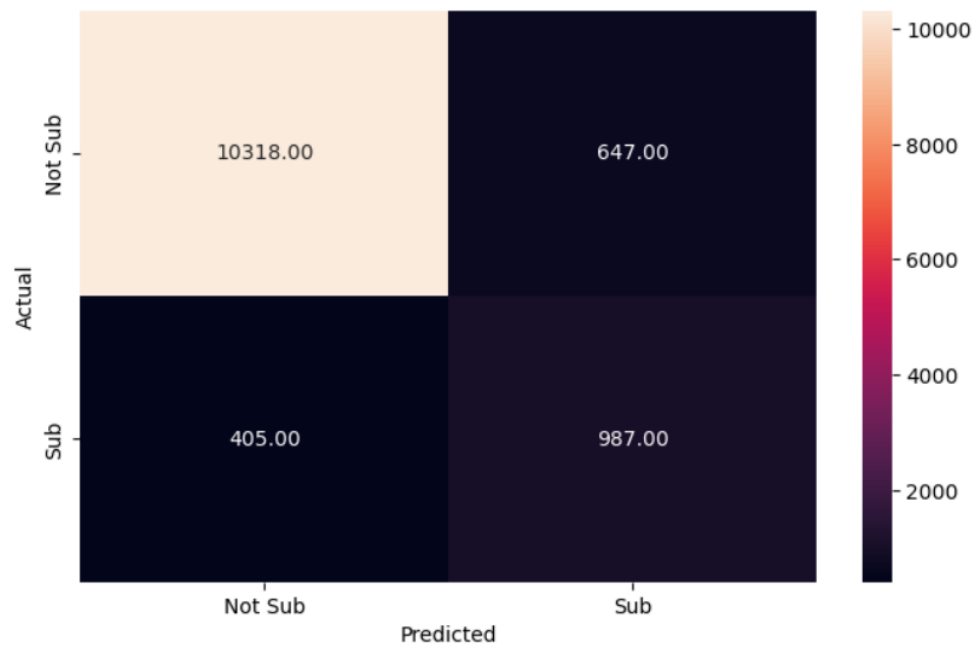
[F6] Random Forest training outcome



[F7] Random forest test data against model



[F8] Hyperparameter tuning on random forest results



Zack Evans
1908686@student.uwtsd.ac.uk

FINAL GRADE

87 / 100

GENERAL COMMENTS

Well Done.

The code is well implemented with comments and explanation of the results. The report is well written, however it could have been improved by adding the theoretical explanation of both models. The report lacks the proper explanation about the evaluation metrics. The presentation of results could have been improved by adding a table to demonstrate and compare the results of both models and the effect of parameter tuning. Need to improve the reference list as per IEEE style. There are some sections in the report that are not completed. Please ensure to proof read the report before submission. Avoid the use of future tense in your report as the report is based on a completed project. Include the figures within the report rather than listing in the annexure as it improves the readability. Try to keep the video presentation is longer than five minutes.

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

GRADING FORM: ACCA7026 GRADING FORM JAN 25

ZACK EVANS

87

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]*



Engages with relevant literature, but lacks depth or synthesis,
Good understanding but lacks theoretical understanding.
Reference style needs to be improved bases on IEEE ref. style

12

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. Clearly explains key hyperparameters, their significance, and impact on model performance with supporting examples. Basic understanding but it lacks adequate explanation.

7

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)



Provides in-depth analysis of model performance using relevant metrics. However this could be improved by adding tables to demonstrate the model performance and the effect of tuning. Significant factors are identified for both models and their impacts are explained. Discusses strengths and limitations but lacks depth or supporting examples

22

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)



Applies appropriate preprocessing techniques, explaining their impact on data quality and model performance.
Implements both models correctly, applies them effectively to the dataset, and provides well-documented results.
Implements models efficiently, applies proper hyperparameter tuning and performance are assessed. Cross validation ?

38

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)



The report is mostly well-organized but lacks clarity in some sections. The figures should be part of the explanation of results rather than listed in the annexure.

8

The code is well-structured, well-commented, and integrates clear visualizations (e.g., plots, confusion matrix) to support analysis.