**LEAD SCORING ANALYSIS USING MACHINE LEARNING**

**PANKAJ YADAV (13578147)**

Student of Engineering, Environment and Computing Department, Coventry University

MSc Data Science (EECT109)

Coventry, United Kingdom

yadavp11@uni.coventry.ac.uk

**Abstract:**

**Lead scoring is an important component of sales and marketing strategy that seeks to discover and prioritise new consumers based on their propensity to convert. Predictive models in machine learning play an important part in automating this process by analysing numerous data points and assigning scores to leads. This study dives into the field of forecasting lead scoring with machine learning techniques. The study investigates the relevance of lead scoring and the benefits of machine learning approaches. This research intends to give insights into the usefulness and accuracy of various algorithms such as logistic regression and random forest in forecasting lead scores by conducting a thorough examination of these models. This project aims to make important contributions by analysing real-world datasets and conducting experiments.**

**Introduction:**

Lead scoring is an important procedure in sales and marketing that identifies and prioritises potential consumers based on their propensity to convert.

A diagram of a sales funnel

Description automatically generated

Fig 1. Leads Stages. Source: (8 Lead Scoring and Lead Qualification Best Practices, n.d.)

Traditionally, lead scoring was based on manually evaluating and set criteria, which frequently resulted in inefficiencies and mistakes in lead evaluation. However, with improvements in machine learning, organisations may now use predictive modelling approaches to automate and optimise the lead scoring process.

The relevance of this project stems from its potential to transform the lead management methods of enterprises across several sectors. By leveraging predictive analytics, businesses may obtain deeper insights into their consumer base, find hidden possibilities, and modify their marketing and sales tactics appropriately. Furthermore, automation of lead scoring reduces the implicit prejudices and inconsistencies associated with manual review, resulting in more objective and data-driven decision-making.  
In the next sections of this study, we will look at the Literature review of this topic and methods used to forecast lead scores using machine learning techniques. We will go over the stages involved in pre-processing the dataset, selecting, and engineering key features, and evaluating several machine learning techniques. Through empirical study and testing, we want to determine the efficacy of our technique and give insights into the elements that influence lead scoring prediction accuracy. Finally, the goal of this project is to help enhance predictive analytics in the marketing and sales domains, allowing organisations to optimise lead management procedures and achieve higher success in client acquisition and retention.  
  
  
**Literature Review:**

A range of studies have explored the use of lead scoring in business.

(Jadli et al., 2022) demonstrate in their paper about the benefits of using machine learning to automate lead scoring processes, comparing different models’ performance in predicting visitor behaviour on business websites. Also, this paper focuses on replacing traditional scoring system with machine learning algorithm for prospect identification.

Shriram in his paper (Human Verification, 2023) discusses the implementation of a predictive lead scoring solution using machine learning to provide sales and marketing teams with advanced knowledge on potential customers.

These studies collectively highlight the potential of lead scoring in improving business processes and the role of machine learning in enhancing its accuracy and efficiency.

**Dataset Description:**

The Dataset which is used to do this project is taken from (*Leads Dataset*, 2019). This project demonstrates the advantages of machine learning in the automation of lead scoring using predictive modelling. I tried with the "X Education" public dataset, which is commonly utilised for lead prediction.

The specifications of the dataset are:

(1). It contains 9240 rows with 37 features.

(2). This dataset is balanced as the target variable contains 60% of instances as ‘0’ and 40% as ‘1’.

A pie chart with numbers and a few percentages

Description automatically generated

Fig 2: Class Imbalance

(3). Different features of this dataset have missing values which are as follows:

|  |  |
| --- | --- |
| Feature Name | No of missing values |
| Lead Source | 36 |
| TotalVisits | 137 |
| Page Views Per Visit | 137 |
| Last Activity | 103 |
| Country | 2461 |
| Specialization | 1438 |
| How did you hear about X Education | 2207 |
| What is your current occupation | 2690 |
| What matters most to you in  choosing a course | 2709 |
| Tags | 3353 |
| Lead Quality | 4767 |
| Lead Profile | 2709 |
| City | 1420 |
| Asymmetrique Activity Index | 4218 |
| Asymmetrique Profile Index | 4218 |
| Asymmetrique Activity Score | 4218 |
| Asymmetrique Profile Score | 4218 |

Table 1: No of Missing Values

(4). The datatypes of the features are float64, int64, and object.

(5). The Target feature is ‘Converted’ which suggest that a particular lead is converted successfully or not.

(6). There are 30 categorical features and 7 numerical features.

**Dataset Pre-processing:**

Before using machine learning algorithms to forecast lead scoring, the data must be pre-processed to confirm its quality and applicability for modelling. Data pre-processing improves predictive model performance and accuracy by resolving problems including missing values, outliers, and feature scaling.

Firstly, to deal with missing values, I applied standard method i.e. replacing missing values in categorical features with mode of that feature and in numerical feature with median.

There are 5 features in this dataset which have all instances as unique (i.e. feature like “Magazine” have all 9240 instances as 0) or binary having irregular shape (i.e. features like “Newspaper” have 9239 instances as ‘No’ and 1 instance as ‘Yes’). So, I dropped these features due to low(or zero) variance.

Categorical Feature Encoding:

The problem with many of the machine learning algorithm is that either they work on numerical data or if the data is categorical, such as months number (1 to 12), they provide biased results towards 12 as compared to 1. To overcome this problem, we will convert our whole data into binary using one hot encoder. After this we will remove dummy variable to reduce the dimensionality. (Brownlee, 2020)

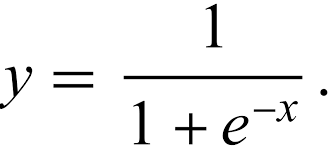
This method results in increasing the number of features from 25 to 104.

**Methodology:**

In this section, I am going to implement three different classification models and will examine their effects, accuracy, precision, recall, and ROC curve. **Scikit-learn** library is mainly used to carry out different model fittings.

***(1). Logistic Regression:***

The use of logistic regression analysis allows for the examination of the connection between (categorical or continuous) independent variable(s) and a single dichotomous dependent variable, unlike linear regression analysis, where the dependent variable is a continuous variable. The discussion of logistic regression in this project is limited in scope. Logistic regression is commonly used for binary classification, which predicts two distinct groups. To do this, the sigmoid function (shown below) is used to compute the outcome and translate numerical results into a probability expression between zero and one.



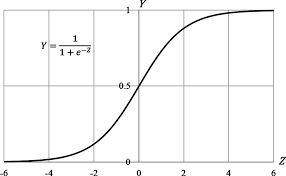


Fig 3:Sigmoid function

Source:(*Plot of The Sigmoid Function.*, n.d.)

Maximum likelihood Estimation(MLE) is used by Logistic Regression to learn about the parameter. when MLE is applied to the Logistic Regression model, it attempts to minimize the below expression:

L (θ)=−​(y​log(pi​)+(1−y)log(1−pi​))

Note that y and pi represent the true and predicted class labels, respectively. The vector θ represents the parameters that the model aims to learn. Therefore, L(θ) measures the model's loss. Gradient descent is a frequent approach for solving MLE's optimisation problems.

***(2). Decision Tree Classifier:***

A decision tree is a supervised learning method used in machine learning to perform classification and regression problems. It offers a visual illustration of a flowchart-like layout containing if-else statements that aids decision-making by forecasting results depending on input factors. Decision trees are popular among machine learning beginners due to their ease of interpretation and comprehension.

Essentially, the mathematics underlying decision trees emphasises optimising the division criteria to construct a tree-like structure that successfully separates the data into homogeneous subsets, allowing for precise forecasts and classifications according to the given data properties. (Fritz, 2023)

***(3). Random Forest Classifier:***

The RF classifier uses a mixture of tree classifiers, each with a random vector sampled randomly from the input data. Each tree ranks for the most popular group to categorise the input vector.

Random forests, like naive Bayese and k-nearest neighbour algorithms, are widely used owing to their ease and high efficiency. Random forests differ from the previous two techniques in that their ultimate model framework is unpredictable. Tree building's stochastic character leads to this outcome. (Caie et al., 2021)

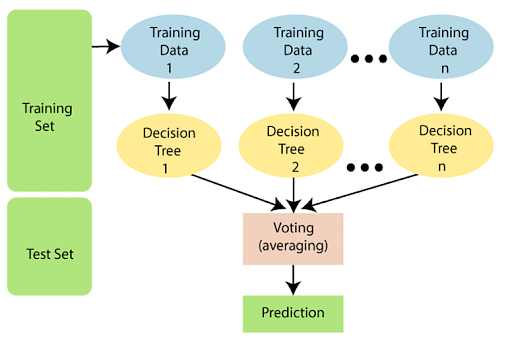


Fig 4: Working of Random Forest Algorithm

Source: (Simplilearn, 2023)

**Results and Discussion:**

The whole python code is run on Jupyter Notebook and standard libraries are used such as pandas and numpy for data importing and pre-processing, scikit- learn for model fitting.

Our conclusions will be based on different parameters:

(1). **Precision**: The frequency with which a machine learning model properly predicts the positive class is measured by a parameter called precision. On dividing the total number of occurrences, the model predicted as positive (including true and false positives) by the number of accurate positive predictions (true positives), you can compute precision. (*Accuracy Vs. Precision Vs. Recall in Machine Learning: What’s the Difference?*, n.d.)

Precision = TP / (TP+FP)

A high value of precision confirms that when a model predicts a positive result, it is likely to be correct.

(2). **Recall**: A machine learning model's recall is a statistic that expresses how frequently it properly selects positive instances, or true positives, out of all the real positive samples in the dataset. Recall may be computed by dividing the total number of positive cases by the number of genuine positives. The latter comprises false negative findings (missed instances) and true positives (successfully discovered cases). (Accuracy Vs. Precision Vs. Recall in Machine Learning: What’s the Difference?, n.d.)

Recall = TP / (TP+FN)

A high recall values shows that the model can effectively identified most of the positive instances in the dataset.

The dataset is divided into 20% as test and 80% as train.

Table 2 provides the concise results of the experiment.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Models | | |
| LR | DT | RF |
| Accuracy (%) | 91.82 | 89.17 | 92.5 |
| Recall (%) | 88 | 81 | 85 |
| Precision (%) | 91 | 91 | 94 |
| F-1 Score (%) | 89 | 86 | 89 |
| ROC curve(AUC) % | 97 | 94 | 97 |
| Training Time (sec) | 0.13 | 0.1 | 0.69 |
|  | Cross Validation | | |
| Accuracy (%) | 88.5 | 88.24 | 91.6 |
| Standard Deviation (%) | 2.6 | 1.6 | 2.3 |

Table 2: Experimental Results

On comparing the results of the experiment, from table 2 Random Forest model is performing good as compared to other two models. It performs with the accuracy of 92.5% whereas logistic regression gives accuracy of 91.82% and Decision tree has 89.17% accuracy.

So, from above we can recommend the use of Random Forest Classifier is good in predicting Leads scoring target feature (i.e. is that customer is converted or not).

**Please note** that the accuracy predicted by Jadli et al., 2022 is more than my accuracy. One of the reasons behind this gap may be the pre-processing of data. In that paper the author got 9074 rows and 89 columns after applying one hot encoder. In my opinion, we can’t delete (/reduce) no of instances of our data. Also, the author deleted the features which have missing values more than 70%. But in my opinion, we can’t drop those features because they may have some valuable information which will affects our final prediction.

The measurement used to assess a classifier's capacity to differentiate between classes is called the Area Under the Curve (AUC).

The ROC curve for above models is represented in Fig 5.

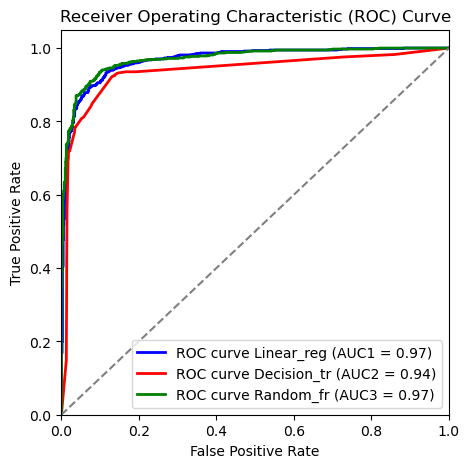


Fig: 5 ROC Curve for different models

**FUTURE WORK:**

In an effort to improve performance, deep learning techniques or neural networks might be used. Additionally, a more thorough grid search strategy in conjunction with piping enables the search for the ideal number of algorithmic components for each of the particular parameters.

**REFERENCES:**

(1)*. 8 Lead scoring and lead qualification best practices*. (n.d.). https://blog.pearagon.com/8-lead-scoring-and-lead-qualification-best-practices

(2). Jadli, A., Hamim, M., Hain, M., & Hasbaoui, A. (2022). TOWARD a SMART LEAD SCORING SYSTEM USING MACHINE LEARNING. *Indian Journal of Computer Science and Engineering*, *13*(2), 433–443. <https://doi.org/10.21817/indjcse/2022/v13i2/221302098>

(3). *Human verification*. (2023). <https://www.semanticscholar.org/paper/Lead-Score-Analysis-Shriram-Tomar/47315cb30e6310544b953aa96d963d27b1ce74aa>

(4).  *Plot of the sigmoid function.* (n.d.). ResearchGate. <https://www.researchgate.net/figure/Fig-41-Plot-of-the-sigmoid-function_fig4_373709457>

(5). Fritz. (2023, September 18). *Understanding the mathematics behind decision trees*. Fritz Ai. <https://fritz.ai/the-mathematics-behind-decision-trees/>

(6). Simplilearn. (2023, November 7). *Random Forest Algorithm*. Simplilearn.com. <https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm>

(7). Caie, P. D., Dimitriou, N., & Arandjelović, O. (2021). Precision medicine in digital pathology via image analysis and machine learning. In *Elsevier eBooks* (pp. 149–173). <https://doi.org/10.1016/b978-0-323-67538-3.00008-7>

(8). *Accuracy vs. precision vs. recall in machine learning: what’s the difference?* (n.d.). Retrieved March 21, 2024, from <https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall>

(9). Brownlee, J. (2020, June 30). *Why One-Hot encode data in machine learning?* MachineLearningMastery.com. <https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>

APPENDIX 1: Python code used to do this experiment is pasted below.

A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

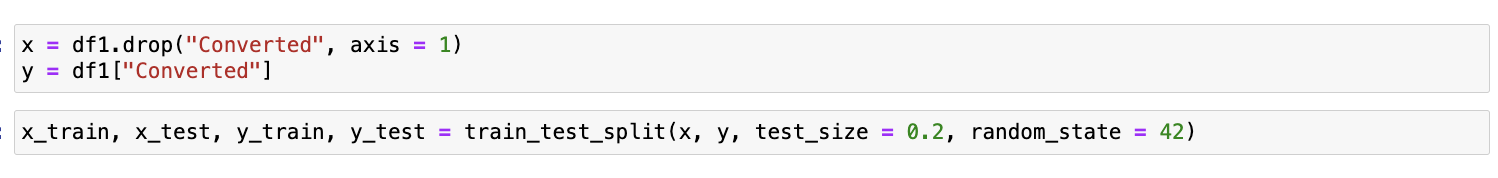
Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer program

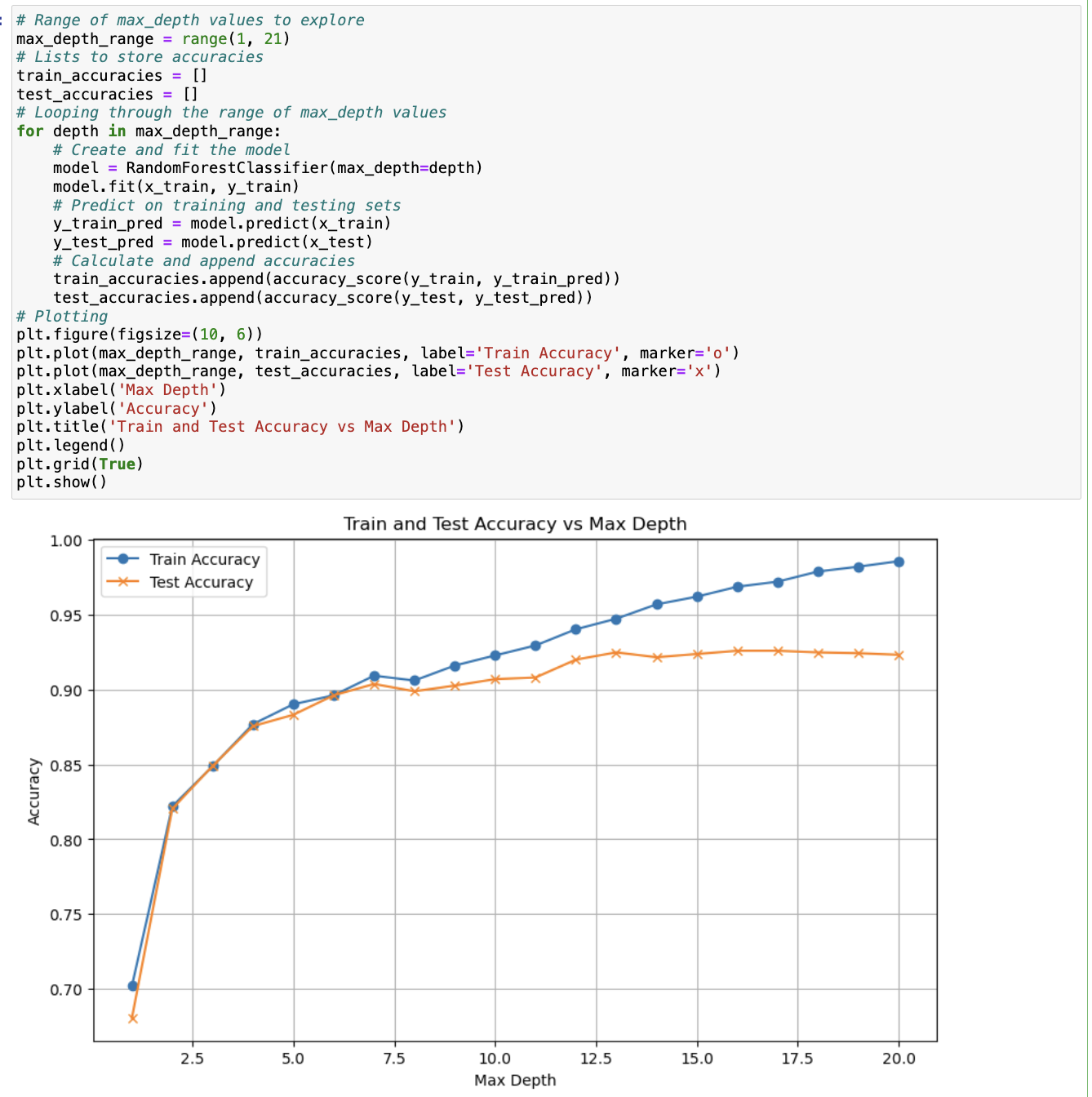
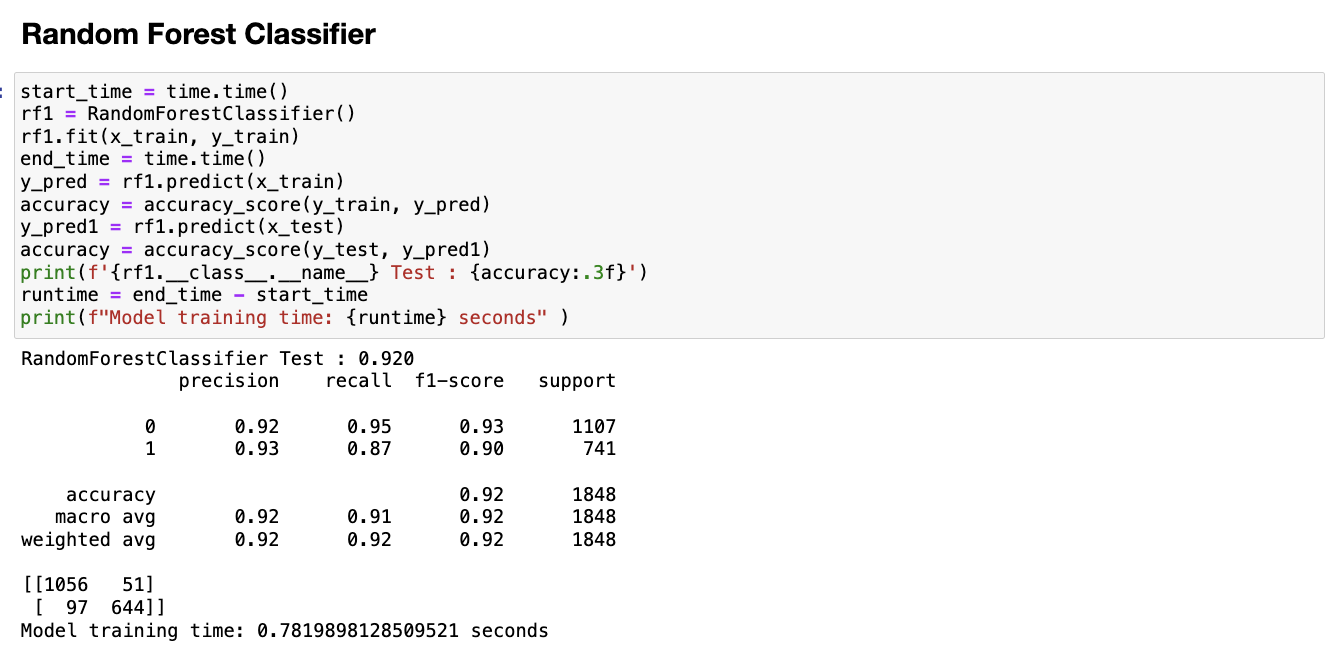
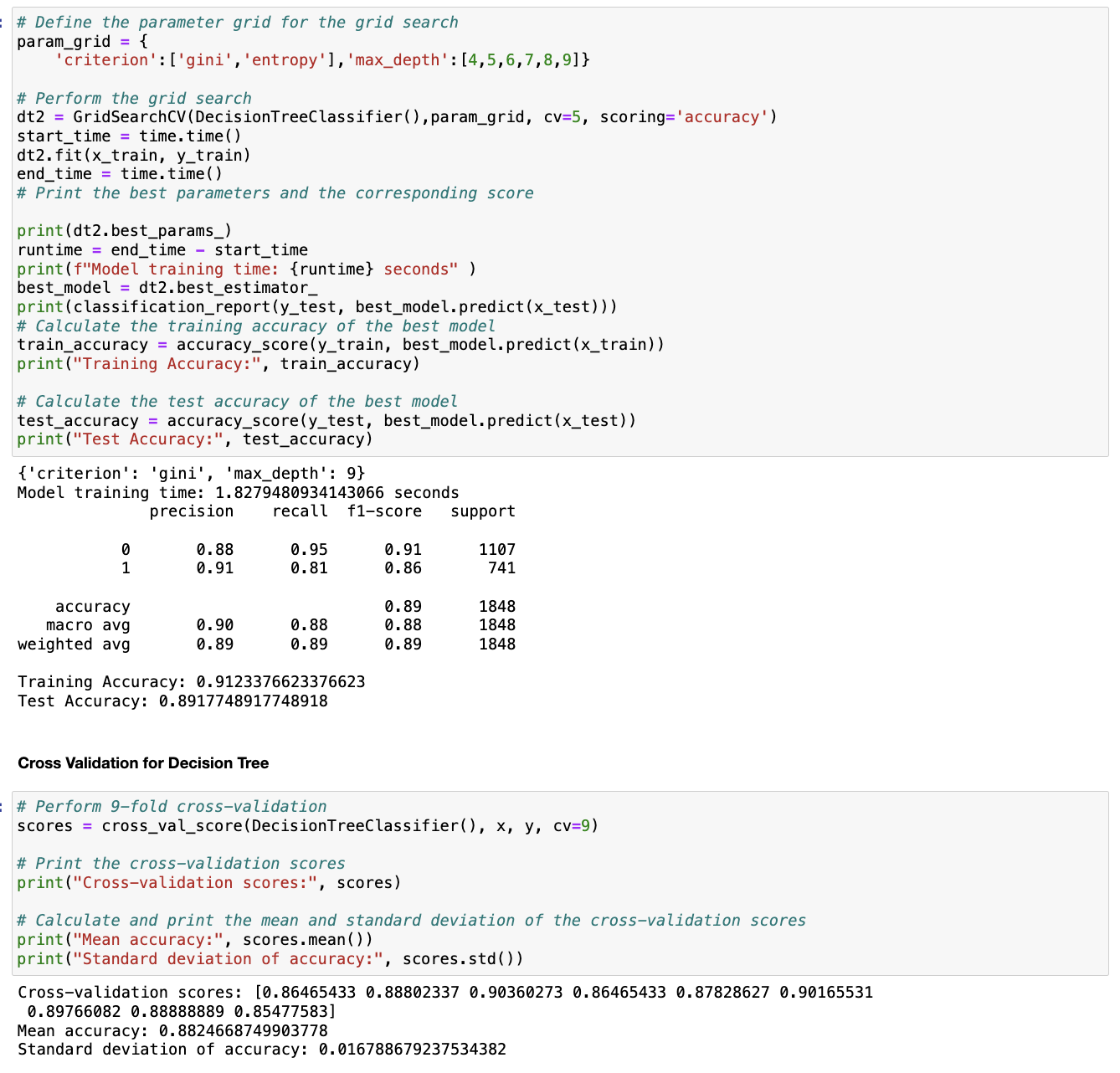
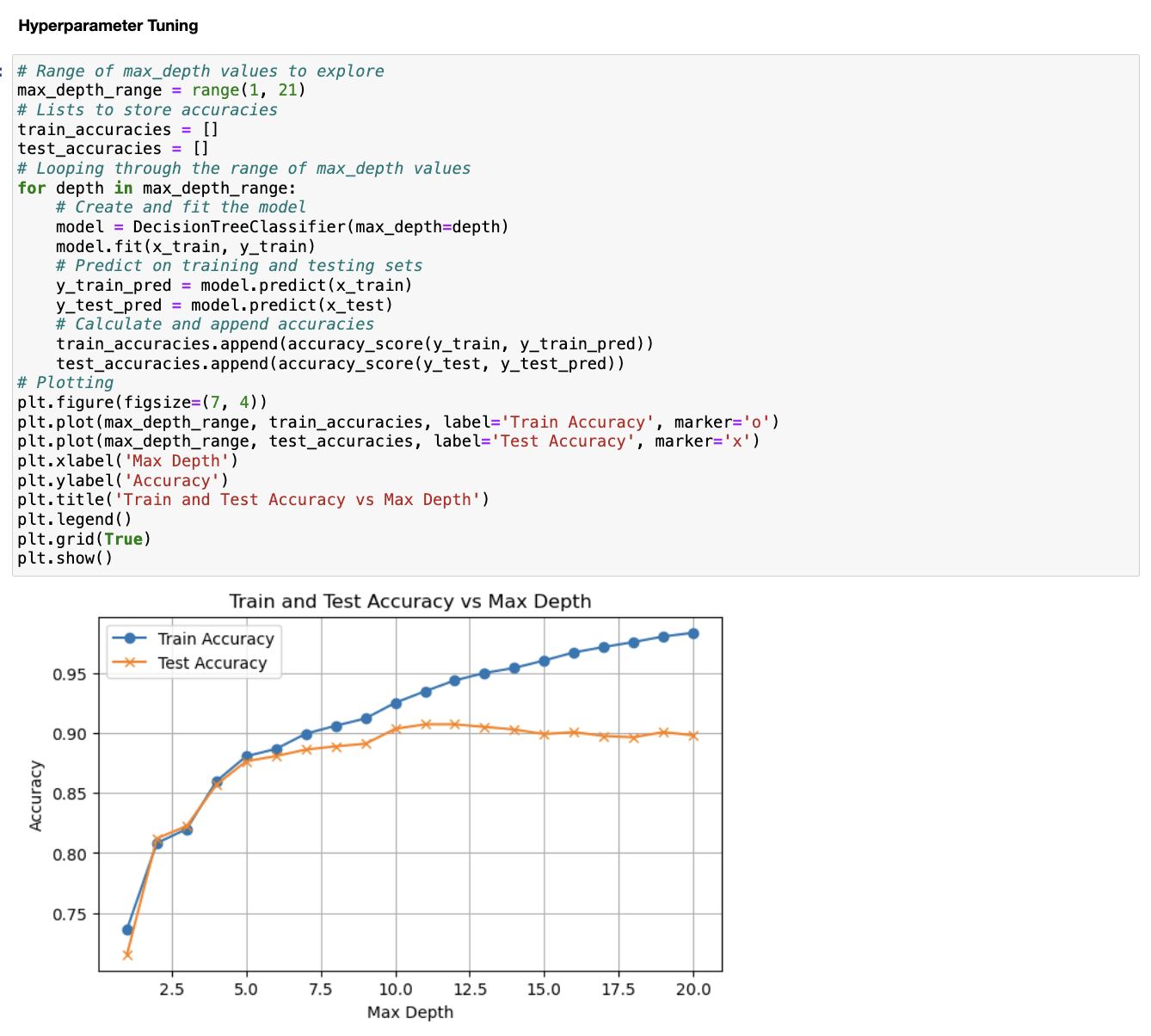
Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generated

A screenshot of a computer program

Description automatically generated



A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated