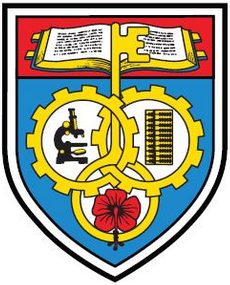
****

**KOLEJ UNIVERSITI TUNKU ABDUL RAHMAN**

**FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY**

**Assignment**

## **BMCS2113 MACHINE LEARNING**

2020/2021

| Student’s name/ ID Number | : | Lee Shu Ern 19WMR0 |
| --- | --- | --- |
| Student’s name/ ID Number | : | Tan Yin Yuen 19WMR05960 |
| Student’s name/ ID Number | : | Hee Sze Wei 19WMR05920 |
| Student’s name/ ID Number | : | Chew Hwa Ern 19WMR04184 |
| Programme | : | RMM2S3 |
| Tutorial Group | : | Group 2 |
| Date of Submission to Tutor | : | 14 May 2021 |

Problem Statement

Term deposits are fixed-term cash investments that an individual deposits into an account at a financial institution for a fixed and agreed maturity period that usually ranges from a few months to a few years and an agreed rate of interest paid into the account. The agreement made for these investments are voided along with an imposed penalty if the individual withdraws their funds before the agreed maturity date. Term deposits are the major source of income for a financial institution, hence the financial institutions usually expand their subscriptions in fixed deposits through various ways such as emails, advertisements in digital marketing or telephonic marketing.

One of the most common ways to market term deposits is through telephonic marketing. Financial institutions tend to invest and focus more on this aspect of marketing to campaign their term deposit products. However, call centers are required to target customers that are most likely to convert their funds to term deposits to save more time and costs in reaching out.

The goal of this study is to determine and classify the clients who are most likely to subscribe or convert funds to a term deposit from the existing client base. This will make it easier for callers to eliminate as many clients who are unlikely to subscribe to a term deposit and focus on reaching out to the clients who are interested.

There have been models made to classify clients for the cross-selling of bank products and services. The result analysis shows that the bank should target customers working in blue collar industries as they have a stable income that does not vary easily. According to the analytics, married customers, clients who are connected through cellular contact and those who possess secondary education or higher, have a rather stronger plausibility in subscribing to other products, while customers who are committed to loans have a poorer prospect. The existing system fit the data into different Machine Learning models and compared the outcome where Gradient Boosting Classifier gave the highest prediction accuracy of 88.79% while Gaussian Naive Bayes gave a rather low prediction accuracy of 86.37%.

The proposed project is used to predict if clients will subscribe to a term deposit in a financial institution. The data used in this project is related to the telemarketing campaign of a Portuguese banking institution.

## The main objective for this dataset:

**Using machine learning techniques to predict the likelihood of financial institution’s clients to subscribe to term deposits through telemarketing campaigns.**

Methodology

• Describes how your solution works on a level that does not lose interesting

details. Make some sensible breakdown into headings (Methodology).

Here is the plan:

1. Data Engineering: check data correctness, fill unknown data cells, mofidy and convert data properties for calculation
2. Exploratory Data Analysis: analyzing data to filter out some main patterns or characteristics
3. Training Models
4. Evaluation

In this proposed project, the data used was preprocessed by checking the data accuracy, eliminating nulled data cells and converting data properties or data type for calculation. The data was split into a training set and testing set to train the models.

The proposed project fits the dataset into several Machine Learning algorithms such as

Logistic regression, KNN, Decision Tree Classifier, Support vector machine, Multinomial Naive Bayes and then compares which model gives the highest accuracy in prediction.

Result

• Includes and explains the evaluation outcome in the article (Result).

##ATTACH APPENDIX 2

Logistic Regression should not be used if the number of observations is less than the number of features; otherwise, it may result in overfitting. It creates linear boundaries. The assumption of linearity between the dependent and independent variables is a major limitation of LogisticRegression. Only discrete functions can be predicted with it. As a result, the discrete number set is bound to the dependent variable of Logistic Regression. Since logistic regression has a linear decision surface, it cannot solve nonlinear problems. In real-world situations, linearly separable data is uncommon.The average or no multicollinearity between independent variables is needed for logistic regression. Complex relationships are difficult to obtain using logistic regression. NeuralNetworks, which are more efficient and compact than this algorithm, can easily outperform it. Theindependent and dependent variables are linearly related in Linear Regression. However, in order to use Logistic Regression, independent variables must be linearly related to the log odds (log(p/(1-p)).

Logistics Regression is most optimally used on the data set chosen in our proposed project as it is the most straightforward to train, apply and interpret. This model provides efficient model training that saves up a lot of time compared to other models such as the Support Vector machine models that may take up to hours to train the model with big datasets. Since our dataset used is considered relatively larger, the Logistic Regression model is more suitable for training the model as it does not require high computational power and It classifies unknown data very quickly. With an imbalanced dataset, it is difficult to obtain the most accurate prediction result. However, The logistic regression model takes into account the predicted parameters by giving inference about the importance of each feature in our dataset to ease the process of finding out significant relationships between each feature.

Moreover, This algorithm allows models to be updated easily to reflect new data, unlike decision trees or support vector machines. The update can be done using stochastic gradient descent or GridSearchCV to train and fit new data into the model.Logistic Regression outputs well-calibrated probabilities along with classification results. This is an advantage over models that only give the final classification as results. If a training example has a 95% probability for a class, and another has a 55% probability for the same class, we get an inference about which training examples are more accurate for the formulated problem.

Overfitting is less common in logistic regression, but it can happen in high-dimensional datasets such as the one used in our proposed project. To prevent over-fitting in these cases, regularisation (L1 and L2) techniques are used. Rather than straight away starting with a complex model, logistic regression is sometimes used as a benchmark model to measure performance, as it is relatively quick and easy to implement. Logistic Regression proves to be very efficient when the dataset has features that are linearly separable. Due to its simple probabilistic interpretation, the training time of logistic regression algorithms comes out to be far less than most complex algorithms, such as an Artificial Neural Network.

On another hand, Logistics regression attempts to predict precise probabilistic outcomes based on independent features. On high dimensional datasets, this may lead to the model being over-fit on the training set, which means overstating the accuracy of predictions on the training set and thus the model may not be able to predict accurate results on the test set. This usually happens in the case when the model is trained on little training data with lots of features. So on high dimensional datasets, regularization techniques are used to avoid over-fitting, where very high regularization factors may lead to the model being under-fit on the training data.

Not only that, Logistic regression has limitations on non-linear problems as it has a linear decision surface. The assumption of linearity between the dependent and independent variables is also a major limitation. Hence prediction accuracy is affected as the data set used in this proposed model is non-linear. To solve that, the transformation of non linear features is required which can be done by increasing the number of features such that the data becomes linearly separable in higher dimensions.

Furthermore, probabilistic predictions made by the model may be incorrect and the model's predictive value may degrade if irrelevant features are used to build the model. Hence, in our proposed model, irrelevant features such as date and month are dropped to improve the accuracy of the prediction outcome. This algorithm is also sensitive to outliers if there are data values that deviate from the expected range in the dataset which may lead to incorrect results.

It is required that each training example be independent of all the other examples in the dataset. If they are related in some way, then the model will try to give more importance to those specific training examples. So, the training data should not come from matched data or repeated measurements. For example, some scientific research techniques rely on multiple observations on the same individuals. This technique can't be used in such cases.