

**J.E. CAIRNES SCHOOL OF BUSINESS & ECONOMICS**

**EXAMINATION SCRIPT COVER PAGE**

|  |  |
| --- | --- |
| **Module Name and Code** |  |
| **Student ID** |  |

In submitting this script, I am aware that it is my responsibility to adhere to the examination guidelines. Please tick (**Yes/No**) for the following:

|  |  |  |
| --- | --- | --- |
|  | **Yes** | **No** |
| I have read the module owners description/expectations for this exam and the submission process. |  |  |
| I have read the examination guidelines. |  |  |
| I am aware of what the University of Galway plagiarism policy entails. |  |  |
| I have submitted a Learning and Educational Needs Summary (LENS) report (if applicable). |  |  |
| I have saved the files for submission (e.g., .docx, .xlsx) **following strictly the format and naming** **required** (e.g., 1187404-MS5107.docx, 1187404-MS5107.docx). |  |  |

**Declaration for this Exam Submission*:***

*In submitting this work I confirm that it is entirely my own. I acknowledge that I may be invited to online interview if there is any concern in relation to the integrity of my exam, and I am aware that any breach will be subject to the University's Procedures for dealing with breaches of Exam Regulations.*

**Table of Contents**

**Section A**

1. Optimisation Modelling
   1. Decision Variables…………………………………………..3
   2. Objective Function…………………………………………..3
   3. Constraints…………………………………………………..4
   4. Answer Report………………………………………………5
   5. Sensitivity Report……………………………………………6
   6. Limits Report………………………………………………...7

**Section B**

1. Data Mining and Predictive Analytics
   1. Business Understanding……………………………………….8
   2. Data Understanding……………………………………………8
   3. Data Preparation……………………………………………….9
      1. Detecting Null and Duplicate values…………..9
      2. Identifying Outliers…………………………….10
      3. Sampling……………………………………….10
      4. Creating Dummies……………………………..10
      5. Data Partitioning……………………………….11
   4. Model Building………………………………………………...11
      1. Neural Networks……………………………….11
      2. Ensemble Boosting……………………………..13
      3. Ensemble Bagging……………………………...14
      4. Radom Trees……………………………............14
      5. Fully Grown…………………………….............15
      6. Best Pruned……………………………..............15
      7. Minimum Error…………………………………15
   5. Evaluation……………………………........................................16
2. Conclusion…………………………………………………………………..16

# **Question 1: Optimisation modelling**

Celtic Candles Inc are manufacturing candles and will supply special holiday candles across **8,500 stores.** They provide **large jar (A), small jar (B), large pillar (C), small pillar (D), and votive candles (E)**. The retailer has agreed to provide at **least 2 feet** for A and C combined**, at least 1.5 feet** for B and D combined, and **at least 1 foot** for E which results in **17,000 feet, 12,750 feet, and 8,500 feet** across 8,500 stores respectively. Celtic Candles Inc currently has **200,000 pounds of wax, 250,000 feet of wick, and 100,000 ounces of holiday fragrance** which are constraints.

Table

Description automatically generated

Our objective is to find out **how many of each product should be made to maximize the profit**

**1.1 Decision Variables:**

A 🡪 Total number of large jars

B 🡪 Total number of small jars

C 🡪 Total number of large pillars

D 🡪 Total number of small pillars

E 🡪 Total number of votive candles

**1.2 The objective function and result variable for the profit formula:**

The objective is to determine the profit incurred by Celtic Candles Inc,

Total profit = (0.25\*A) + (0.20\*B) + (0.24\*C) + (0.21\*D) + (0.16\*E)

**1.3 Uncontrollable variables representing Constraints are given below:**

The given problem contains material and display constraints which are present below,

**Table

Description automatically generated**

The number of jars sold must be great than the number of pillars sold which is also taken as a constraint (Note: Since the solver doesn’t have the greater than option, we have chosen greater than or equal to option)

In excel **SUMPPRODUCT** function has been used and noted under the **‘Used’** column. We will be using **Simplex LP** to determine the maximum profit making sure the constraints are satisfied. For our initial model we have considered A = 100, B = 100, C = 100, D = 100 and E = 100

Table, Excel

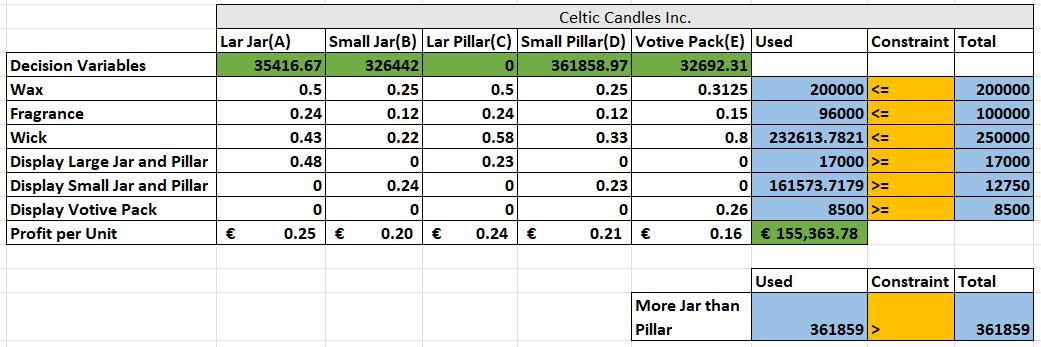
Description automatically generated

The decision variables and profit has been highlighted in green. In the solver, we have added **$H$12** as the objective as we need to determine the profit, **$C$5:$G$5** as the changing variable cells as they are decision variables, and constraints are added as well,

Graphical user interface, text, application

Description automatically generated

We are using Simple linear programming to solve this problem and upon selecting SOLVE option **Answer report, Sensitivity report and limits report** will be generated and the details are displayed below,



**1.4 Answer report:**

A picture containing application

Description automatically generated

The **Objective cell** compares the profit obtained by the company before and after the solver. Initially, with the decision variables as 100 and without adding the constraints we obtained a maximum profit of **€106** but after adding the constraints we got an optimal solution of **€155363.78** which means there is an increase of **€155257.78**,

Table

Description automatically generated

The **Variable cells** contain the original value and final value which represent the decision variable before and after the solver respectively. Initially, we considered 100 of each candle to be produced and based on constraints the solver produced the optimal solution. It produced Votive pack = 32692, Jars = 361858, and Pillars = 361858. The number of jars produced is equal to the number of pillars produced in order to get maximum profit and it is advised not to produce any large pillars,

Table

Description automatically generated

From the **Constraints table**, the slack indicates zero if it is completely utilized which shows **binding nature** but there is some value in slack for wick and fragrance showing **non-binding nature** which means it is not utilized to its maximum limit. So, the solver suggests these materials can be used later on by the management for a different purpose. But for the small jar and pillar, 161553 ft is needed which means each store needs to allocate 19ft to attain maximum profit.

**1.5 Sensitivity report:**

Table

Description automatically generated

From the **Variable cells,** we can deduce that the maximum profit can obtain when zero large pillars are produced and production of large jar = 35416, small jar = 326442, small pillars = 361858, and votive pack = 32692. Reduced cost of large pillar indicates a value of -0.1 which means if the objective coefficient is increased by 0.1 which represents 0.24 – (-0.1) = 0.34 then we can produce at least 1 large pillar. The allowable increase and decrease represent that the objective coefficient can be changed in one of two ways which mean it can be increased or decreased without altering the decision variables.

Table

Description automatically generated

From the **Constraints table,** the final value and R.H side stand for used and available material/space. We can see the shadow price and it represents how much the final value changes if we change the R.H. side based on a given value. For example, the shadow price of Wax is 0.82 which means there will be a profit of €0.82/lb where the allowable increase = 8333 lb which means an additional 8333lb of wax can be used, therefore the total profit = 0.82\*8333 which is €6833.

**1.6 Limits report:**

The **limits report** shows another form of sensitivity report and the overall profit incurred after the solver produced the result based on the decision variables and constraints,

Graphical user interface, text, application

Description automatically generated with medium confidence

Diagram

Description automatically generated with medium confidence

The decision variables Large Jar, Small jar, Large Pillar, Small Pillar, and Votive Pack = 35416, 326442, 0, 361858, and 32692 respectively which produced the optimal solution with an overall profit of €155363. The lower limit and upper limits are varying based on the production of the candles.

# **Question 2: Data mining and predictive analytics.**

We will do research and evaluate the bank dataset in order to categorize the customers who will be potential clients for City Commerce bank. To gather the perfect model for the prediction we will utilize CRISP-DM (Cross-Industry Standard Process for Data Mining), which involves the below steps before classifying the data,

Diagram

Description automatically generated

## **2.1 Business Understanding**

In order to increase the efficiency with which it sells term deposits to prospective customers, City Commerce Bank's management is constantly striving to enhance its marketing tactics. The objective is to create a model for the bank to comprehend its clientele and forecast client reactions based on its telemarketing campaign in order to target those clients with its marketing initiatives. We will evaluate the dataset using the provided data, create models, and select the best model based on prediction accuracy and minimal error in order to categorize the future client.

The best model chosen based on the accuracy will be recommended to the City Commerce Bank for deployment.

## **2.2 Data Understanding**

The Bank dataset contains **4522** records including a header that consists of three types of variables which is binary, numerical, and categorical variables. The dataset contains information about previous customers where each of the columns represents details like age, job, and marital status of the customer and also has a few columns on how the customers responded during the telemarketing campaign.

All the columns and the meanings are explained below,

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Column Name | Type | Meaning |
| Input Variables | Age | Numeric | Age of the Customer |
| Job | Categorical | Type of job |
| Martial | Categorical | Marital Status |
| Education | Categorical | Education of the Customer |
| Default | Binary | Has Credit in default? |
| Balance | Numeric | Average yearly balance |
| Housing | Binary | Has a housing loan? |
| Loan | Binary | Has a personal loan? |
| Contact | Categorical | Contact communication type |
| Day | Numeric | Last contact day of the month |
| Month | Categorical | Last contact month of the year |
| Duration | Numeric | Last contact duration |
| Campaign | Numeric | number of contacts performed during this campaign and for this client |
| Pdays | Numeric | number of days that passed by after the client was last contacted from a previous campaign |
| Previous | Numeric | number of contacts performed before this campaign and for this client |
| Poutcome | Categorical | outcome of the previous marketing campaign |
| Output Variable | y | Binary | Has the client subscribed to a term deposit? |

## **2.3 Data Preparation**

We will pre-process the data to fulfil the standards by cleaning, constructing, integrating, transforming, and rescaling the dataset before we dive into the actual classification model.

**2.3.1 Detecting Null and duplicate values:**

From the dataset, we were not able to find any NULL or duplicate values,

A screenshot of a computer

Description automatically generated

**2.3.2 Identifying Outliers:**

An outlier is a data point in statistics that dramatically deviates from other observations. The following formula is used to compute the mean and standard deviation for each column of numerical data and to determine whether any of the numbers are outliers

**[mean - 3\*stdev; mean + 3\*stdev]**

In excel we will use a custom formula using IF and OR for more understanding,

=IF(OR(F2<=$F$4524-3\*$F$4525,F2>=$F$4524+3\*$F$4525),"OUTLIER","NO")

From the dataset, we were able to figure out **88** outliers for Duration, **87** outliers for Campaign, and **99** outliers for Previous. We decided not to discard the outliers when assessing this because they are not statistically different, which implies that they are not more than three times the original values and are not too much outside of the current range.

**2.3.3 Sampling:**

The dataset contains a total of **4521** records and since the number exceeds 1000 we will need to perform sampling in order to work on samples from the entire population. Since the output variable is in binary, we have chosen **Stratified random sampling** so that we get equal responses of ‘yes’ and ‘no’ from the dataset which resulted in **1042** records.

Graphical user interface, text, application, email

Description automatically generated

**2.3.4 Variable handling - Creating dummies:**

Data mining cannot be used for model classification with categorical variables. In order to resolve this, we will be creating dummies for **Job, Marital, Education, Default, Housing, Loan, Contact, Month, Poutcome, and Y.**

Transform 🡪 Transform categorical data 🡪 Create dummies

Graphical user interface, application

Description automatically generated

**2.3.5 Data Partitioning:**

Before data modeling, the data is partitioned for each variable and also a couple of variables are removed in order to avoid redundancy.

The partitions values contain,

1. Training – Contains training partition of **625** records
2. Validation – Contains validation partition of **417** records

**Partition Summary**

Graphical user interface, application

Description automatically generatedGraphical user interface

Description automatically generated with medium confidence

## **2.4 Model Building**

Identifying customers who are interested in choosing term deposits is the key goal. In the dataset obtained from the telemarketing campaign, we are examining the client profiles. We will be utilizing the below methods and choose the best to figure out the potential customers based on the maximum accuracy and minimum error,

1. Neural Networks
2. Ensemble Classification
3. Classification Tree

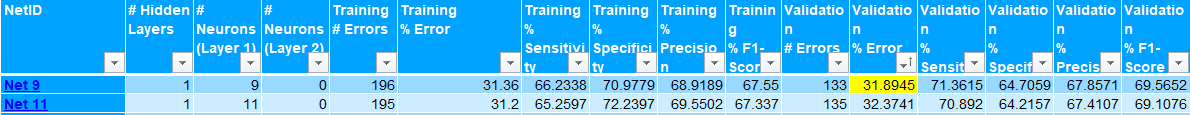
**2.4.1 Neural Networks:**

Neural Networks is a machine learning algorithm that receives a huge amount of data and develops a function to classify the information. It is mostly made up of several ordered layers of neurons.

We are determining the set with the best minimum error for the classification model,

Classify 🡪 Neural Networks 🡪 Automatic Network

We are using Normalization as rescale data and changing the Epoch to figure out the best method. By using an Epoch value of **500** we get a validation error of **31.89** whereas we got **32.85, 34.77, 38.12, and 38.84** for epoch values for **300, 100, 30, and 20** respectively.



Graphical user interface, application

Description automatically generated

We get the minimum error for **Layer 1 with 9 Neurons** and the same has been used for the classification model below with **Epoch = 500**,

Classify 🡪 Neural Networks 🡪Manual Network

Graphical user interface, application

Description automatically generated

A picture containing treemap chart

Description automatically generatedTable

Description automatically generated

Using Neural networks, we got an **Overall Error = 31.89% and Accuracy = 68.10%.**

**Ensemble Classification:**

Ensemble seeks to improve the machine learning results by combining a couple of weaker models to form a strong model. There are three types,

1. Ensemble Boosting
2. Ensemble Bagging
3. Random Trees

Under weak learners, we have a couple of methods like Decision Tree, Neural Network, Logistic Regression, k-Nearest Neighbors but we will be using Decision tree to run all the classification models.

**2.4.2 Ensemble Boosting:**

In order to create a strong model, Boosting uses weak models in conjunction with incorrectly classified records. And Boosting's method builds tactics that improve future prediction by drawing on previous blunders.

Classify 🡪 Ensemble 🡪 Boosting

When we run the model for the STDPartition we get **CBoosting\_Output, CBoosting\_TrainingScore, and CBoosting\_ValidationScore** so we will concentrate on the 3rd sheet to confirm if this model is suitable for City Commerce Bank to classify its customers.

Table

Description automatically generated with medium confidenceTable

Description automatically generated

From the above, we can infer the overall **error percentage is 19.90% which means its 80.09% percent accurate and has a precision value of 0.79**

**2.4.3 Ensemble Bagging:**

It is also called Bootstrap aggregation. From the original dataset, it creates numerous training datasets using random sampling. All operations in Boosting take place sequentially whereas in Bagging all steps take place in parallel.

Classify 🡪 Ensemble 🡪 Bagging

We get the same results compared to Boosting but when we look at the values, we can infer that there is a decrease in the error value in turn the accuracy increases.

Table

Description automatically generated with medium confidenceTable

Description automatically generated

Bagging uses multiple decision trees simultaneously.

For Bagging, we got an **Error percentage = 18.94% and Accuracy = 81.05% with a precision of 0.79** which means it’s a better classification model compared to Boosting technique.

**2.4.4 Random Tree:**

The Random tree is classified under Bagging and uses multiple separate decision trees and it’s considered to be more accurate.

Classify 🡪 Ensemble 🡪 Random Tree

When looking at the validation score sheet we can see the overall error percentage is more and accuracy is less compared to both Boosting and Bagging.

Table

Description automatically generated with low confidence

Table

Description automatically generated

The random tree has an **error value = 23.26% and accuracy = 76.73% with a precision of 0.77** which suggests that it’s not a suitable model.

Hence comparing all three models we can conclude is the best classification model in **Ensemble Bagging with an Accuracy = 81.05**

**Classification Tree:**

When developing a model, the classification process is carried out using a decision tree, also known as a classification tree. Data must be partitioned and then further divided iteratively on each branch. There are three types to look at and we will choose the best model based on the error and accuracy,

* + 1. Fully grown
    2. Best pruned
    3. Minimum error

Classify 🡪 Classification tree

And under the decision tree, we need to select Prune using the validation set. Pruning in data mining usually eliminates the leaves and branches to increase the overall performance so, we can run the model for all three types sequentially.

1. For **Fully grown** we got Error = 19.66% and Accuracy = 80.33% with a precision of 0.80
2. For **Best pruned** we got Error = 20.86% and Accuracy = 79.13% with a precision of 0.81
3. For **Minimum Error** we got Error = 18.94% and Accuracy = 81.05% with a precision of 0.80

Based on the above analysis, we can conclude that the minimum error can be considered the suitable model in the classification tree to classify the customers who will opt for term loans

**Full Grown - Error report and Metrics from the ValidationScore**

Table

Description automatically generated Treemap chart

Description automatically generated

**Best pruned - Error report and Metrics from the ValidationScore**

Table

Description automatically generated Treemap chart

Description automatically generated with medium confidence

**Minimum Error - Error report and Metrics from the ValidationScore**

Table

Description automatically generated Chart, treemap chart

Description automatically generated

## **2.5 Evaluation**

Let’s take an overall look at the below table with all the model outputs,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | % Error | % Accuracy | Precision | F1 Score |
| Neural Network | 31.89 | 68.1 | 0.67 | 0.69 |
| Boosting | 19.9 | 80.09 | 0.79 | 0.8 |
| Bagging | 18.94 | 81.05 | 0.79 | 0.82 |
| Random Tree | 23.26 | 76.73 | 0.77 | 0.77 |
| Fully Grown | 19.66 | 80.33 | 0.8 | 0.8 |
| Best Pruned | 20.86 | 79.13 | 0.81 | 0.79 |
| Minimum Error | 18.94 | 81.05 | 0.8 | 0.81 |

By comparing all the models, we can immediately rule out the neural network, random tree, and best pruned because they have very high error rates and poor accuracy. While the values of Boosting, Bagging, fully grown, and Minimum error are nearly equivalent when compared and despite the values being close, we have chosen **Minimum error** from the classification tree as the best model to handle City Commerce Bank's problem of classifying customers who will choose a term deposit based on telemarketing.

Table

Description automatically generatedTable

Description automatically generated

Table

Description automatically generated

From the confusion matrix we determine the below values,

1. **Accuracy(#correct)** = ‘0’ Predicated cases/ ‘1’ Predicted cases = 162+176 =338
2. **Accuracy(%correct)** = Number of Actual predictions/Validation records = 338/417 = 81.05%
3. **Precision =** True Positive / (True Positive + False Positive) = 176/ (176+42) = 0.8
4. **Sensitivity =** True Positive/ (True Positive + False Negative) = 176/(176+37) = 0.82
5. **F1 Score =** 2 \*((Precision\*Sensitivity)/(Precision+Sensitivity)) = 2\*((0.8\*0.82)/(0.8+0.82)) = 0.81

Even though the accuracy and error numbers for bagging and minimum error are equal we have chosen minimum error as the best model for City Commerce Bank since it has a greater precision of **0.80**.

Since all the values are very close, we have also analyzed the Lift chart and ROC curve.

From the lift chart, we can see the average line (red) and predicted values line (blue) are fairly separated with a gap from each other hence the model performed well with the selected data for minimum error.

From the ROC curve, we can see the ROC value = **0.80524** which is close to 1 means minimum error is a better model compared to others also since the curve is close to the top left, we can confirm the model has performed well in classification.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

## **3. Conclusion**

The Minimum error classification model created is stored in the sheet named **MinError\_Stored** and during the final stage of CRSIP-DM model called **Deployment**, the model can be deployed where the customer data is further analyzed and monitored in order to classify the customers and also address new open issues.