A close-up of a logo

Description automatically generated

**Examination Assignment Cover Page**

|  |  |
| --- | --- |
| **Module Code** | **MS5107** |
| **Student name and ID** | **Snigdha Chaudhuri**  **23102262** |

In submitting this script, I am aware that it is my responsibility to adhere to the examination guidelines. Please tick (double click->… or Y/N) for the following:

|  |  |  |
| --- | --- | --- |
|  | **Yes** | **No** |
| I have read the module owners description/expectations for this exam assignment and the submission process. |  |  |
| I have read the exam assignment guidelines. |  |  |
| I am aware of what the University of Galway plagiarism policy entails. |  |  |
| I have saved the files for submission (e.g., .docx, .xlsx) **following strictly the format and naming** **required** (e.g., **12345678\_MS5107\_ExamAssignment.docx, 12345678\_MS5107\_ExamAssignment\_Q1.xlsx,**  **12345678\_MS5107\_ExamAssignment\_Q2.xlsx**). |  |  |

**Declaration for this Assignment Submission*:***

*I hereby declare that the work submitted is entirely my own work. It has not been taken from the work of others, except to the extent that such work has been cited and acknowledged within the text of my work. This work is not done in whole or in part by a machine or through Generative Artificial Intelligence, such as ChatGPT or else. I have not and will not allow anyone to copy my work thereby helping a peer to complete their assignment.*

**Table of Contents:**

|  |  |  |
| --- | --- | --- |
| Sno | Description | Page No. |
| 1 | Question 1: Optimisation modelling | 3 |
| 2 | Question 2: Data mining and predictive analytics. | 7 |

**Table of Figures:**

|  |  |  |
| --- | --- | --- |
| **Sno.** | **Items** | **Page No.** |
| 1 | Fig 1: Image of the model with the constraints | 3 |
| 2 | Fig 2: Image of the Solver Dialogue box | 4 |
| 3 | Fig 3: Image of the Answer Report generated after the model was run. | 4 |
| 4 | Fig 4: Image of the Sensitivity Report generated after the model was run. | 5 |
| 5 | Fig 5: Image of the Limits Report generated after the model was run. | 6 |
| 6 | Fig 6: Image showing the imputation report. | 8 |
| 7 | Fig7: Image of creating an encoded dataset of selected variables. | 8 |
| 8 | Fig 8: Image of data set partitioning. | 9 |
| 9 | Fig 9: Image for the reports generated for Logistical regression | 10 |
| 10 | Fig 10: Image for the reports generated for Ensemble Random Trees | 12 |

**Question 1: Optimisation Modelling**

**Model:**

A table with numbers and numbers

Description automatically generated

Fig 1: Image of the model with the constraints

The above model is developed to find the quantity of products to maximize the profitability of 8500 stores.

Model Formulation:

The below-mentioned steps were used to run the simplex LP method to solve the above question:

**Step 1: Define Decision Variables.**

The following are the decision variables-

Quantities of-

1. Lange Jar
2. Small Jar
3. Large Pillar
4. Small Pillar
5. Votive Pack

**Step 2: Formulate the Objective Function to maximize profit-**  
 Maximize Z(Profitability)=0.25\* (Large Jars)​+0.20\*(Small Jars)​+0.24\*(Large Pillars) +0.21\*(Small Pillars)​+0.16(Votive Pack)

**Step 3: Identify the constraints:**

* Total wax usage < 200000 lbs
* Total Quantity of Fragrance < 100000 oz
* Total Wick < 100000 feet
* Large Jars and Large pillars occupy at least 2 feet per shop.
* Small Jars and small cups occupy at least 1.5 feet per shop.
* Votive Pack occupies at least 1 feet.
* Jars should be greater than Pillars

**Step 4: Run the Model:**

**Solver Dialogue Box-**

A screenshot of a computer

Description automatically generated

Fig 2: Image of the Solver Dialogue box

The model is constructed to get the maximum profit by adjusting the manufactured products' quantities. After running the Simplex LP model, the following reports were generated-

1. **Answer Report:**

A screenshot of a data

Description automatically generated

Fig 3: Image of the Answer Report generated after the model was run.  
  
The findings from the Solver offer a valuable glimpse into a winning production strategy for Celtic Candles Inc, geared towards maximizing profit. Let's delve into the insights that illuminate the optimization of production quantities and the strategic considerations involved:

**Profit Maximization:** The refined production plan promises a total profit of €154,693.0361. This signifies the most lucrative blend of product quantities within the given constraints.

**Optimal Product Mix:**

* Large Jars: 35,416.67 units
* Small Jars: 225,968.82 units
* Large Pillars: 0 units
* Small Pillars: 441,899.77 units
* Votive Packs: 49,038.46 units

The recommendation here is to skip producing Large Pillars, directing resources toward more profitable products.

1. **Sensitivity Report:**

**Variable Analysis:**

* **Reduced Cost:** Large Jars and Small Jars show zero reduced costs, signaling an ideal solution. The negative reduced cost for Large Pillars hints that producing more could boost overall profit.
* **Shadow Price Analysis:** Shadow prices give us insight into how profit reacts to changes in constraint values. Adjustments in constraints related to displays, wax, fragrance, wick, and product quantities can have distinct impacts on profitability.
* **Constraints Analysis:** Constraints linked to display sizes for Large Jars and Large Pillars, Small Jars and Small Pillars, and Votive Packs are considered binding which means that these are the constraints that currently limit expansion, and increasing display sizes could enhance overall profits.

A screenshot of a computer

Description automatically generated

Fig 4: Image of the Sensitivity Report generated after the model was run.

1. **Limits Report:**

A screenshot of a computer

Description automatically generated

Fig 5: Image of the Limits Report generated after the model was run.

The above report image gives us the following details-

* **Profit Stability:** The company can count on making around €154,693.0361 if they stick to the plan, if they adjust the product quantities within the given ranges.
* **Flexibility in Production:** The company has room to change the quantities they make without losing out on profits. This helps them adapt to things like changes in customer demand or availability of resources.
* **No Need for Large Pillars:** The plan doesn't include making Large Pillars, but if they want to, they can make a tiny bit without affecting the profits much.

**Recommendations:**

* Display Adjustment: Fine-tune display sizes to explore opportunities for increased profits. The model hints that expanding displays could positively impact overall profitability.
* Fragrance Optimization: Investigate ways to maximize fragrance usage within the allocated budget. The positive shadow price for fragrance suggests the potential for increased investment in this area to boost overall profit.
* Flexibility Monitoring: Keep a close eye on constraints with "Binding" slack. While currently acting as limits, periodic reassessment might reveal opportunities for adjustments, providing more flexibility in production planning.

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

**Introduction:**

The main objective of this report is to understand, analyze, and train the model created with the data set provided, to predict whether a customer in a bank will invest in a term deposit or not.

The data set provided to us includes variables with all the details of the customer interaction with the bank including their age, type of job, marital status, education type, credit default, account balance, current loans with the bank, type of communication with the bank, with the date and duration. The dataset also provided campaign-related communication or contact with the customers. The provided dataset will help in providing a classification model to predict the probability of customers opting for a term deposit.

**CRISP-DM Methodology:**

**Business Understanding:**

Western Alliance Bank is currently dealing with challenges in optimizing its marketing strategies, especially in persuading customers to subscribe to term deposits. The existing telemarketing campaigns are falling short in achieving the desired subscription rates, requiring a more targeted and data-driven approach.

Goals:

* Boost Term Deposit Subscription Rates: The primary aim is to increase the number of customers opting for term deposits.
* Strategic Marketing Optimization: Develop a predictive model to enhance the efficiency of marketing efforts, ensuring resources are directed towards customers more inclined to subscribe.
* In-Depth Customer Understanding: Gain profound insights into the customer base, enabling the bank to craft personalized and effective marketing plans.

Context:

Telemarketing campaigns involve reaching out to potential customers via phone calls to promote term deposit subscriptions. The success of these campaigns hinges on identifying key factors that influence customer decisions and leveraging this information for more personalized interactions.

**Data Understanding:**

**Dataset Overview:**

The 'Bank.xlsx' dataset contains information from past telemarketing campaigns, covering various customer attributes and outcomes. Notable variables include age, job type, marital status, education, credit default, balance, housing, and loan status, contact history, and the outcome of the previous marketing campaign.

**Relevant Variables:**

Demographics: Age, job, marital status, and education provide insights into the customer's background.

Financial Status: Balance, housing, loan, and credit default offer information about the customer's financial health and obligations.

Communication History: Contact type, day, month, and duration describe the details of the last contact, providing context for campaign effectiveness.

Campaign History: Number of contacts, days since the last contact (pdays), and outcome of the previous campaign (poutcome) offer insights into the customer's engagement history.

**Data Preparation:**

The data was initially checked for missing data, which came out to be zero.Below is the screenshot for the imputation sheet which shows 0 records deleted.-

A screenshot of a computer

Description automatically generated

Fig 6: Image showing the imputation report.

The next step is to remove any outliers that may be present in the data to remove the error that may occur due to high variability in the data. Here we removed the outliers which was 12 percent of the data using the formula: (mean-/+ 3\* standard deviation) for the numeric data set provided for the variables, age, duration, and balance.

Post the outlier removal, the categorical variables with non-numeric were encoded and converted into One-hot coding. The below image shows the number of records and Columns that were changed to numeric coding. -

A screenshot of a computer

Description automatically generated Fig7: Image of creating an encoded dataset of selected variables.

The next step is to use the standard partitioning method to divide the dataset into training and validation sets for us to train the model and validate it post the training.

A screenshot of a computer

Description automatically generated

Fig 8: Image of data set partitioning.

The data is now set to be utilized for training and validation using different models.

**Modelling:-**

1. **Model Analysis-**

XLMiner was used to build the data model. The below-mentioned models were explored-

1. Neural networks
2. Ensemble- Bagging, Boosting, Random Trees
3. Logistic Regression
4. Discriminant analysis
5. K-nearest neighbors

The model was trained using the training data set and validated using the validation dataset.

1. **Model Description-**

Logistic Regression is used in predictive analytics, particularly when faced with binary classification tasks - situations where we need to decide between two outcomes, like saying "yes" or "no" to a specific scenario. To break it down:

* Sigmoid Function (Sigmoid): Logistic Regression employs a special mathematical function called the Sigmoid. This function takes various inputs, such as a person's age, job type, and whether they have a term deposit, and transforms them into a probability value between 0 and 1.
* Score Aggregation: We assign scores to different features, essentially giving importance to factors like age, job type, and others. These scores are summed up, resulting in a collective or "secret" score.
* Probability Estimation: The Sigmoid function then transforms this cumulative score into a probability, signifying the likelihood of someone saying "yes" to a term deposit. If the probability surpasses 0.5, we predict a "yes"; otherwise, we predict a "no."
* Model Training: To enhance its predictive prowess, the model undergoes training. It's exposed to a myriad of examples where we already know the outcomes, and it learns from these instances to refine its predictive abilities.
* Iterative Learning: Like a diligent learner, the model continually adjusts itself based on feedback. It aims to minimize discrepancies between its predictions and the actual outcomes.
* Feature Significance: The scores assigned to different features indicate their significance in influencing the model's predictions. Positive scores imply a higher likelihood of a positive outcome, while negative scores suggest the opposite.
* Probabilistic Communication: Unlike delivering binary decisions, the model communicates in probabilities. For instance, it might express confidence by saying, "I am 70% certain this person will say 'yes'." We can then apply a threshold, such as 50%, to make a final decision.

In essence, Logistic Regression acts as a sophisticated analytical companion, diligently examining examples, learning from them, and offering nuanced predictions about whether someone will express interest in a term deposit. It involves a touch of mathematical finesse, making it a robust and valuable tool for making informed decisions.

1. **Evaluation:**

Presenting the metrics for 2 best model out of which Logistical Regression was chosen.

**Model 1 Analysis:**

**A screenshot of a graph

Description automatically generated A blue and black text

Description automatically generated**

Fig 9: Image for the reports generated for Logistic regression.

1. **Confusion Matrix:**

Here the ‘yes represents the customer investing in fixed deposit and ‘no for not.

* True Positives (1,1): 60 - These are instances where we predicted "yes" (1) correctly.
* True Negatives (0,0): 1403 - These are instances where we predicted "no" (0) correctly.
* False Positives (0,1): 30 - We predicted "yes" when it was actually "no."
* False Negatives (1,0): 99 - We predicted "no" when it was actually "yes."

1. **Error Report:**

* The error rate for saying "no" (Class 0) is low at 2.09%.
* The error rate for saying "yes" (Class 1) is high at 62.26%.
* The overall error rate across both classes is 8.10%.

1. **Metrics:**

* Accuracy: We got 91.90% of predictions correct overall.
* Specificity: We're really good at saying "no" when it's "no" (97.91%).
* Sensitivity (Recall): We're okay at saying "yes" when it's "yes" (37.74%).
* Precision: When we say "yes," we're right about 66.67% of the time.
* F1 score: A good balance between saying "yes" and "no."
* Success Class and Probability: We care more about predicting "yes" (Class 1). We set our threshold for predicting "yes" at 0.5.

1. **Explanation:**

We've been using a model to predict if customers will sign up for something (let's call it "yes") or not ("no").

* **How Good Are We?:**
* Overall, we're about 92% accurate, which is pretty good.
* When we say "no," we're almost always right (98% of the time).
* But when we say "yes," we're not as good (38% of the time).
* **Mistakes We Make:**

Sometimes, we mistakenly say "yes" when it's really "no" (2% of the time for "no", which is the Type 1 error). The bigger issue is when we miss saying "yes" when it's actually "yes" (62% of the time for "yes", which is a Type 2 error).

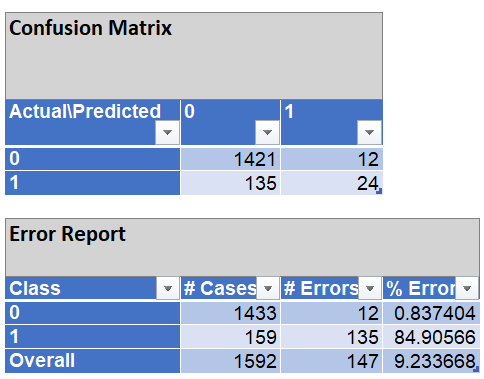
* **What the Numbers Mean:**

We've got some good points: We're cautious when saying "yes," but we need to get better at not missing out on actual "yes" cases.

* **How We Can Improve:**

We might need to adjust our model a bit to catch more of those "yes" moments without making too many mistakes.

**Model 2 Analysis**:

 A blue and black text with white text

Description automatically generated

Fig 10: Image for the reports generated for Ensemble Random Trees

1. **Confusion Matrix:**

* True Positives (1,1): The model correctly predicted that 60 customers would sign up for term deposits, and they did.
* True Negatives (0,0): The model accurately said 1403 customers wouldn't sign up for term deposits, and they didn't.
* False Positives (0,1): In 30 cases, the model said customers would sign up, but they didn't.
* False Negatives (1,0): In 99 cases, the model said customers wouldn't sign up, but they did.

1. **Error Report:**

The error rate for Class 0 is low at 0.84%, but for Class 1, it's high at 84.91%.

The overall error rate across both classes is 9.23%.

1. **Metrics:**

* Accuracy: The model correctly predicted 90.77% of cases.
* Specificity: The ability to correctly identify the negative class (0) is high at 99.16%.
* Sensitivity (Recall): The ability to correctly identify the positive class (1) is very low at 15%.
* Precision: Of the instances predicted as positive, 66.67% were positive.
* F1 score: A balance between precision and recall, indicating a moderate performance.

1. **Success Class and Probability:**

The model designates Class 1 as the success class. The success probability threshold is set at 0.5.

The confusion matrix and associated metrics suggest that while the model exhibits high accuracy, it struggles particularly with identifying positive cases (Class 1). The low sensitivity indicates a high number of false negatives, impacting the model's ability to catch positive instances. Precision is moderate, indicating room for improvement in avoiding false positives. Further optimization may be required to enhance the model's performance, particularly in capturing positive outcomes.

1. **Deployment:**

Deploy the model for use in predicting whether a customer will subscribe to a term deposit.

**Model Proposal:**

**Justification for Model Choice:**

Logistic Regression was the chosen model for the following reasons-

Strengths:

* Pretty Accurate: Overall accuracy is high at 91.9%.
* Good at Saying "No": It's great at correctly identifying when something is not happening (0).
* Okay Precision: When it predicts something positive, it's right about 38% of the time.

Areas to Work On:

* Needs Improvement for Saying "Yes": It's not as good at identifying positive cases, as shown by the low sensitivity.

The classification model is a suitable choice for binary classification problems, such as predicting term deposit subscriptions.

**Additional Insights:**

**1**. Identifying Key Features Influencing Subscription Decisions:

* In looking at our data and model, some factors stand out as key influencers on whether customers sign up for term deposits:
* Time Spent Talking (duration): The longer a customer talks during the last call, the more likely they are to subscribe. Engaging in conversations seems to make a difference.
* Previous Campaign Results (poutcome): If a customer responded positively in the past, they might likely do so again. Learning from what worked before can guide future efforts.
* Balance: A customer's average yearly balance is crucial. Those with higher balances seem more inclined to subscribe, suggesting financial stability plays a role.
* Contact Method (contact): How we get in touch matters. Whether it's by phone or cell, different channels can affect a customer's decision to sign up.
* Age: Different age groups might have different views on term deposits. Tailoring our approach based on age can make our marketing more effective.

2. Recommendations for Targeted Marketing Based on Customer Profiles:

* Prioritize Engaging Conversations: Focus on making conversations during telemarketing longer and more interesting. The more engaged a customer is, the more likely they are to say yes.
* Learn from Success: Take cues from what worked in previous campaigns. If a strategy led to positive outcomes, use it again. Learning from mistakes helps refine our approach.
* Segment by Balance: Split customers into groups based on their average yearly balance. Customize messages to address the financial needs of different segments.
* Optimize Communication Channels: Figure out which communication channels work best. Whether it's a phone call or a text, use the channels that customers respond to most.
* Age-Tailored Messaging: Craft messages that suit different age groups. Younger customers may care about different things than older ones.
* Educate Customers: Make sure our marketing materials are informative. Clearly explain why term deposits are a good idea and address any concerns. Informed customers are more likely to sign up.

By focusing on these insights and recommendations, Western Alliance Bank can boost the impact of its marketing efforts.

**Conclusion:**

In conclusion, the proposed Logistic Regression model offers a balance between simplicity and effectiveness for predicting term deposit subscriptions. Its interpretability and performance metrics make it a valuable tool for the bank's marketing strategy. Additionally, insights gained from the model can guide future marketing campaigns towards more targeted and successful efforts.