Milestone 4: Data Modeling

Group D

CONTENTS

INTRODUCTION

In this milestone, we aim to build upon the initial model developed in Milestone 3 by refining and finalizing the model. The goal is to ensure the final model accurately predicts customer eligibility for satellite internet service based on socio-economic factors, such as marital status, education, occupation, annual salary, gross year-to-date earnings, household size, and years of residence as selected.

The process involves selecting a modeling technique, cleaning, and processing the data (including converting categorical variables into numeric format), splitting the dataset for training and testing, building the model, and evaluating its performance based on predefined success criteria. This document outlines the modeling process, the technique used, the assumptions made, and the final model evaluation.

TARGET VARIABLE

The target variable **eligibility** was designed to classify customers as either eligible or ineligible for satellite internet service. The eligibility was based on specific financial and socio-economic criteria related to income and stability.

CRITERIA FOR ELIGIBILITY

- **Annual Salary > R50,000**:
  - Financial Stability: Customers with yearly incomes over R50,000 are well-off and have steady incomes, which enables them to pay for the service.
  - The affordability threshold serves as a fundamental affordability filter, guaranteeing that the customers under consideration earn enough money to cover the service without experiencing financial hardship.

- **Years of Residence > 3 years**:
  - Customer Stability: If a customer has been residing at their home for more than three years, it suggests that they are stable and dependable, which may be related to their financial status and future capacity to continue making service payments.
  - Reduced Risk: Because these people are less likely to default or move often, the organization has less risk.

- **Gross Year-To-Date Earnings > R45,000**:
  - Current Profits Insight: A more recent picture of a customer's financial status may be obtained by looking at year-to-date earnings. A criterion for gross year-to-date profits guarantees the customer's financial stability at the time of evaluation.
  - Enhanced Eligibility Confidence: By establishing this cutoff, customers are further filtered to those who have maintained a sufficient income throughout the year.

MODELLING TECHNIQUE SELECTION

SELECTED TECHNIQUE: RANDOM FOREST

Random Forest was chosen for the final model due to its robustness, ability to handle many features (both categorical and numerical), and effectiveness in reducing overfitting. It is an ensemble learning method that constructs multiple decision trees during training and merges them to improve accuracy and stability.

WHY RANDOM FOREST?

- **Versatility with Data Types:** The dataset includes categorical variables (such as marital status, education, and occupation) and numerical variables (such as annual salary and gross year-to-date earnings). Random Forest handles both types of data with ease.
- **Non-linear Relationships:** Random Forest excels at capturing non-linear relationships within the data, which may exist between variables like income and household size, or occupation and years of residence.
- **Handling High Dimensionality:** The algorithm efficiently handles many features and selects the most relevant ones during the model-building process.
- **Feature Importance:** Random Forest can identify the most influential variables, which helps in understanding key drivers behind customer eligibility for Satellite services. This feature is crucial for business insight, as it allows stakeholders to focus on the most relevant factors.

CHALLENGES ADDRESSED BY RANDOM FOREST

- **Overfitting:** Individual decision trees can overfit the data. Random Forest mitigates this risk by averaging the predictions of multiple trees, providing a more generalized model.
- **Class Imbalance:** If there is a significant imbalance between eligible and non-eligible customers, Random Forest can still perform well by using techniques like class weighting and balancing its bootstrapped samples.

DATA PREPARATION

The dataset required careful cleaning and preprocessing before model training. We started by removing irrelevant columns and transforming categorical variables into a numeric format. This step was critical to ensuring compatibility with the Random Forest algorithm and improving model accuracy.

CATEGORICAL DATA TO NUMERIC CONVERSION

To prepare the data, we converted key categorical features into numeric values:

- **Marital Status:** The categories (Divorced, Married, Single, Widow, Unknown) were assigned numeric values ranging from 1 to 5.

- **Education:** Education levels (Bachelors, HS-grad, Masters) were mapped to numeric values from 1 to 3.

- **Occupation:** The occupation categories (Cleric, Executive, Professional, Sales) were converted to numeric values (1–4).

This transformation was necessary because Random Forest algorithms require numerical input to calculate distances and construct decision splits. Without this step, the algorithm would treat categorical variables incorrectly or inefficiently.

REMOVAL OF IRRELEVANT COLUMNS

We removed irrelevant or redundant columns to focus on the most meaningful predictors. Specifically, year_of_birth and age_group were eliminated as they either duplicated information or were not directly tied to the customer's eligibility. Additionally, columns like Gross_Pay_Last_Paycheck and Gross_FRS_Contribution were excluded to avoid multicollinearity and focus on variables most relevant to the business problem.

TEST DESIGN

To evaluate the Random Forest model's performance accurately, the dataset was split into training and testing sets. This separation ensures that the model is trained on one portion of the data and then tested on an unseen portion, which allows for a more accurate assessment of its predictive ability.

DATA SPLIT

- **Training Set (75%):** Most of the dataset was used for model training, allowing the Random Forest algorithm to learn patterns and relationships within the data.
- **Testing Set (25%):** A smaller portion of the data was set aside for testing the model's performance on unseen data, ensuring that the model is not overfitting and can generalize well to new data points.

EVALUATION METRICS

- **Accuracy:** A measure of how many predictions the model got right compared to the total number of predictions.
- **Confusion Matrix:** This metric provides insight into the number of true positives, false positives, true negatives, and false negatives, which is particularly useful when the classes are imbalanced.
- **Precision and Recall:** These metrics were computed to evaluate the model's ability to correctly identify eligible and non-eligible customers. Precision focuses on the accuracy of positive predictions, while recall emphasizes the model's ability to capture all actual positives.
- **F1-Score:** The harmonic mean of precision and recall, providing a single measure of model performance that balances both metrics.

BUILDING THE MODEL

The Random Forest model was trained using the training data. Several model parameters were carefully chosen and tuned to balance accuracy with computational efficiency.

KEY MODEL PARAMETERS

- **Number of Trees (ntree = 10):** A small number of decision trees was used to prevent overfitting while still providing high accuracy. Increasing the number of trees would improve accuracy at the cost of computation time, so 10 trees offered a balanced trade-off.
- **Number of Variables Tried at Each Split (mtry):** The number of predictors considered at each split was automatically set to the square root of the total number of predictors. This is a standard approach in Random Forests, helping to reduce variance while keeping bias low.

MODEL BEHAVIOR AND INTERPRETATION

Accuracy: The model showed high accuracy on the training set, correctly classifying a significant majority of the customers as eligible or non-eligible for satellite services.

Robustness: The Random Forest algorithm is robust to outliers and noise in the dataset making it an excellent choice for this type of problem.

Feature Importance: Variables such as years of residence, annual salary, and household size were identified as the most important predictors in the model. This is consistent with domain knowledge, where customers with stable incomes and long-term residency are more likely to afford satellite services.

MODEL EVALUATION

After training the Random Forest model, we evaluated its performance on the testing data. Various metrics were computed to assess how well the model generalized to unseen data, ensuring that it could be reliably deployed in practice.

CONFUSION MATRIX AND PERFORMANCE METRICS

Accuracy: The model achieved high accuracy reflecting its strong predictive power in classifying eligible and non-eligible customers.

Confusion Matrix: The confusion matrix provided insights into the model’s true positives (correctly classified eligible customers), false positives (incorrectly classified non-eligible customers), true negatives, and false negatives.

Precision, Recall, and F1-Score: The model’s precision and recall were evaluated to ensure it could reliably classify both classes. The F1-Score balanced these two metrics, confirming that the model performed well across both precision and recall.

ROC Curve: The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were also used to further evaluate the model's discriminative ability between eligible and non-eligible customers.

CONCLUSION AND MODEL INTERPRETATION

The Random Forest model proved to be an effective tool for predicting customer eligibility for Satellite services based on socio-economic data. The model's accuracy and robustness were high, and key features such as years of residence, annual salary, and household size emerged as important predictors.

Key Insights:

- **Affordability Indicators**: Customers with higher salaries, longer years of residence, and smaller household sizes were more likely to be classified as eligible for Satellite services.
- **Predictive Power**: The model was able to generalize well to new data, with minimal overfitting. This makes it a reliable tool for deployment in business scenarios.
- **Feature Importance**: The insight into feature importance can be leveraged to prioritize which socio-economic factors to focus on when targeting potential customers or designing service offerings.