**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission.*

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | CA2 project |
| **Lecturer Name:** | Muhammad Iqbal |
| **Student Full Name:** | Muqaddas Mehmood |
| **Student Number:** | 2023485 |
| **Assessment Due Date:** | 15/12/2023 |
| **Date of Submission:** | 15/12/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Project Report: Titanic Dataset Analysis**

**1. Introduction:**

The analysis of the Titanic dataset encompasses data processing, feature engineering, and the application of a Random Forest classification model. The report succinctly outlines the key steps in each phase, introduces the chosen methodology cantered around the Random Forest algorithm, and presents results emphasizing accuracy and discriminatory power via the ROC curve and AUC value. This narrative serves as a concise yet comprehensive exploration of the machine learning journey, from data inception to insightful conclusions.

**2. Data Processing and Feature Engineering:**

**2.1 Dataset Overview:**

The Titanic dataset, recognized for its historical significance and conveniently accessible on Kaggle, unfolds like a captivating tapestry, offering an all-encompassing perspective on the passengers of the iconic Titanic voyage. With a substantial structure of 1309 rows and 28 columns, each data point meticulously contributes to the portrayal of an individual's journey. It's akin to a carefully crafted painting, where each row and column acts as a deliberate brushstroke, intricately detailing the narrative of life aboard this legendary vessel. This extensive dataset, comprised of 28 distinct features, adds layers of complexity and depth, enriching the nuanced representation of passengers and their diverse experiences.

***Survived****:* This binary variable, indicating whether a passenger survived (1) or not (0), serves as the pivotal target variable for the ensuing classification task.

***Pclass***: Denoting the passenger's ticket class (1st, 2nd, or 3rd), Pclass unveils socio-economic disparities among passengers, offering a lens into the ship's diverse demographics.

***Name***: More than just a nominal identifier, the 'Name' feature presents an opportunity for extracting familial relationships, social status, or cultural background.

***Sex***: Categorizing passengers into male or female, 'Sex' reflects the gender composition of the Titanic's passengers.

***Age***: As a numerical feature, 'Age' provides insights into the age distribution aboard the ship, contributing vital demographic context.

***SibSp and Parch***: These features quantify familial relationships, encapsulating the count of siblings/spouses and parents/children on board, respectively.

***Ticket, Fare, Cabin***: Illuminating details about the ticket, fare paid, and cabin number, these features enrich our understanding of passengers' accommodations.

***Embarked***: Indicating the port of embarkation—Cherbourg ('C'), Queenstown ('Q'), and Southampton ('S')—this feature adds a geographic dimension to the dataset.

**2.2 Exploratory Data Analysis (EDA):**

We started looking closely at the data with a process called Exploratory Data Analysis (EDA). Here are some of the steps we followed.

***Handling Zero Columns:*** Rigorous scrutiny revealed a set of columns riddled with zeros, devoid of substantial information. The decision to drop these columns stemmed from their lack of correlation with the target variable, thereby refining the dataset for meaningful analysis.

***Mitigating Missing Values:*** The 'Embarked' column, exhibiting missing values, underwent meticulous imputation, replacing the zeros with contextually relevant data. This strategy ensured the dataset's completeness.

***Data Types and Missing Values***: The `info()` function emerged as a key ally, unveiling the dataset's structure, highlighting data types, and flagging areas with missing values. This function served as a compass, guiding subsequent data processing endeavours.

***Numerical Summaries***: Delving into numerical summaries using the `describe()` method, statistical measures like mean, standard deviation, and quartiles were unearthed. These statistics facilitated the identification of outliers and offered insights into the distribution of numerical variables.

***Visualizations:*** Beyond numerical insights, the narrative of the dataset unfolded through compelling visualizations. Bar charts, box plots, and histograms wove a visual tapestry, elucidating the age distribution and offering an immediate glimpse into the frequency of classes within the target feature. An example of a bar chart showcasing the frequency of classes in the target feature is presented below:A graph of a distribution of survival

Description automatically generated

This thorough initial exploration laid the groundwork for subsequent data processing steps and informed decisions on feature engineering and missing data handling. Understanding the dataset's nuances and distributions was critical for the subsequent machine learning analysis. This intricate dance between data processing and exploratory analysis served as the prologue to subsequent chapters of modelling and evaluation, ensuring that the Titanic dataset was not just a historical artifact but a rich canvas ready for machine learning analysis.

**3. Methodology:**

**3.1 Selection of Machine Learning Algorithm:**

The methodology phase stands as a pivotal juncture where the choice of a machine learning algorithm shapes the trajectory of the analysis. For the Titanic dataset, the discerning selection fell upon the Random Forest classification algorithm, a decision rooted in its adaptability to diverse data structures and its proficiency in handling nuanced relationships within the dataset.

***Random Forest Overview:***

At its core, a Random Forest constitutes an ensemble learning method that assembles a multitude of decision trees during training. The final output is determined by the mode of the classes predicted by individual trees, making it a formidable tool for classification tasks. The distinguishing attributes of Random Forests include their prowess in capturing intricate relationships, accommodating both numerical and categorical features, and mitigating overfitting through the amalgamation of multiple trees.

***Versatility and Flexibility:***

Random Forests exhibit remarkable versatility, requiring minimal hyperparameter tuning. Their capacity to handle a mix of feature types, coupled with their resilience to overfitting, positions them as a robust choice in machine learning endeavours.

***Handling Complex Relationships:***

The Titanic dataset harbours features with potentially nonlinear relationships. The inherent strength of Random Forests lies in their adeptness at capturing these complexities, rendering them particularly apt for discerning patterns within the dataset.

***Effective for Binary Classification:***

Given the nature of the task—predicting the survival status of passengers (binary classification)—Random Forests emerge as a fitting choice. Their inherent ability to efficiently discriminate between the two classes, coupled with their capacity to unveil feature importance, solidifies their suitability for this specific scenario.

In summation, the meticulous consideration given to the selection of the Random Forest algorithm underscores the methodology's commitment to aligning the chosen model with the dataset's characteristics and the overarching objectives of the analysis. The versatility and efficacy of Random Forests lay a robust foundation for subsequent phases of model evaluation and analysis.

**4. Model Analysis and Evaluation:**

**4.1 Model Evaluation Metrics:**

The rigorous evaluation of the Random Forest model employed a set of meticulous metrics to gauge its predictive prowess, ensuring a comprehensive understanding of its performance dynamics.

***Accuracy:*** Serving as a foundational metric, accuracy delves into the overall correctness of the model. It encapsulates the proportion of accurately predicted instances, encompassing both true positives and true negatives. While providing a high-level view of correctness, accuracy may be nuanced in scenarios of imbalanced datasets.

***Confusion Matrix:*** Offering a granular breakdown of predictions, the confusion matrix dissects the model's performance into four quadrants—true positives, true negatives, false positives, and false negatives. Each quadrant unveils specific insights, delineating the model's proficiency in discerning positive and negative instances.

***Classification Report:*** Elevating the evaluation beyond accuracy, the classification report provides a holistic view of precision, recall, and F1-score for each class. Precision articulates the accuracy of positive predictions, recall (sensitivity) gauges the model's ability to capture all positive instances, and the F1-score strikes a balance between precision and recall. This nuanced trio of metrics offers a more refined perspective, especially crucial in scenarios where the consequences of false positives or false negatives vary.

In conclusion, the analysis transcends mere correctness and delves into the intricacies of the model's performance across different facets. This multifaceted evaluation strategy enhances the interpretability of results and facilitates more informed decisions concerning the model's utility and potential areas for refinement.

**4.2 ROC Curve:**

The incorporation of the Receiver Operating Characteristic (ROC) curve elevates the model analysis, offering an additional layer of insight into the Random Forest model's discriminatory power across varying probability thresholds. The ROC curve serves as a dynamic visual representation, plotting the true positive rate against the false positive rate and unveiling the model's discriminative prowess.

**Purpose of Evaluation Metrics:**

***Accuracy:*** While serving as a foundational metric, accuracy may fall short in providing a nuanced understanding in the context of imbalanced datasets. It provides a high-level overview of overall model correctness.

***Confusion Matrix:*** Beyond accuracy, the confusion matrix dives into the specifics of the model's performance, delineating strengths and weaknesses. It dissects predictions into true positives, true negatives, false positives, and false negatives, offering a more granular perspective.

***Classification Report:*** This multifaceted report delves into precision, recall, and F1-score for each class. Precision measures the accuracy of positive predictions, recall gauges the model's ability to capture all positive instances, and the F1-score strikes a balance between precision and recall. This nuanced trio provides a more refined perspective, crucial in scenarios with varying consequences for false positives and false negatives.

The holistic integration of these evaluation metrics collectively crafts a comprehensive narrative of the Random Forest model's predictive capabilities. Accuracy and the confusion matrix unveil the model's proficiency in correctly classifying survival outcomes, while the classification report adds granularity with precision and recall insights. The ROC curve, with its true positive rate versus false positive rate dynamics, provides a nuanced understanding of the model's performance across different decision thresholds.

**5. Results and Conclusions**

The findings stemming from the evaluation of the Random Forest model on the Titanic dataset paint a picture of promising efficacy, with an attained accuracy of 88%. This accuracy metric encapsulates the proportion of correctly predicted instances within the entire dataset, providing a foundational overview of the model's correctness.

Complementing accuracy, the ROC curve steps into the limelight as a pivotal player in assessing the model's discriminatory capabilities. This visual representation delineates the trade-off between the true positive rate and false positive rate across diverse classification thresholds. The ROC curve's utility is further underscored by the derivation of an Area Under the Curve (AUC) value, a numerical indicator of discriminatory performance. In this analysis, the ROC curve yielded an AUC value of 92%, signifying a robust ability to distinguish between positive and negative classes.

The ROC curve, with its AUC companion, delivers a nuanced and comprehensive assessment of the Random Forest model's predictive prowess. Its portrayal of the model's performance across various decision thresholds adds depth to the analysis, enhancing the understanding of how the model responds to different levels of stringency in classification.

A graph of a curve

Description automatically generated

The graphical representation of the ROC curve, as showcased above, visually encapsulates the model's aptitude in discriminating between survival and non-survival classes. This visualization serves as a powerful tool for stakeholders and analysts, offering a tangible depiction of the model's strengths in handling classification tasks.

In the grand finale, the Random Forest model emerges as a stalwart performer on the Titanic dataset. The commendable accuracy, coupled with the robust AUC value, positions the model as an effective tool for binary classification tasks. These results lay a sturdy foundation for future refinements and applications in subsequent stages of the analysis, underscoring the Random Forest model's significance in unravelling insights from the Titanic dataset.

**6. References:**

1. McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference.

2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

3. Heptapod. (n.d.). Titanic Dataset. Retrieved from Kaggle: <https://www.kaggle.com/datasets/heptapod/titanic>

4. Github Repository Link: <https://github.com/Muqaddasmehmood/strategic-thinking-CA2.git>