**Project Title**

**Predictive Modeling for Effective Customer Response Classification**

**Introduction**

**Identifying the Problem:**

In the banking environment, it is significant for organizations to maintain as well as attract customers (Sağlam & Montaser, 2021). The dataset chosen for this project, bank-additional-full.csv sourced from *https://archive.ics.uci.edu/dataset/222/bank+marketing*, concerns the results of a telemarketing campaign by a bank. Specifically, it seeks to estimate the likelihood that a certain client will take up a term deposit by adopting machine learning analysis of the demographic information, behavioral patterns, as well as economic activity levels. The problem is vital for banking as efficient Customer Response Modelling helps identify the appropriate channels giving an accurate reflection of the bank’s customers’ reactions towards certain adverts; thus, helping in the efficient utilisation of resources in a bid to enhance the success rates when conducting marketing campaigns.

**Primary Objectives:**

1. Prediction Accuracy: Develop a machine learning model that will effectively classify whether or not a customer will subscribe to term deposit (employing a binary classification, subscribe or not).

2. Feature Importance: Determine the dominant antecedents that enable the customers to decide on the subscription.

3. Actionable Insights: Give directions on how telemarketing promotions can be improved for best results in the future.

4. Explainability: Make it feasible for the model to generate conclusions easily explainable by computer and more important by human beings.

The goal of this project is to bring together and classify real-world dataset using machine learning approaches as well as feature engineering.

**Training Supervision**

**Model Learning Process:**  
The type of learning for this particular problem is the supervised learning type because the type of data provided here is the training data labeled data also the target variable y, denoting if the customer subscribed for term deposit or not is supported by the research study done by Rony et al. (2021). The model shall be trained in order to be able to generalize some factors that stand as a probability to subscription.

**Task Type:**  
This is a binary classification task, where the goal is to classify customers into two categories: there will only be two types of people: those to answer “yes” and those to answer “no”. This problem is nicely amenable to classification algorithms such as Logistic Regression, Decision Trees, Random Forests, or methods using Boosting (such as XGBoost).

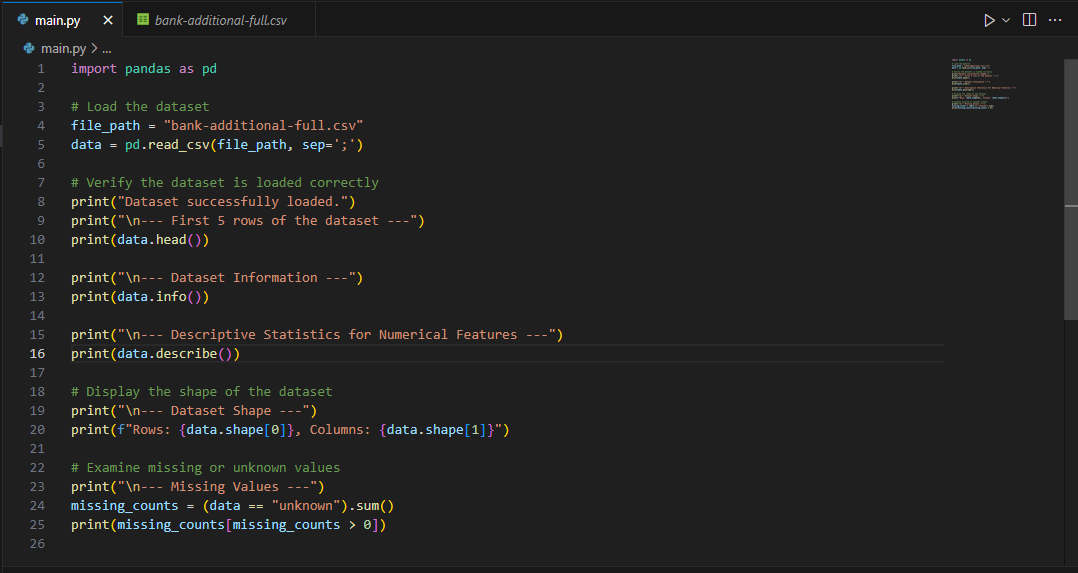
**Learning Technique:**  
The model will use batch learning in which the model is trained using the entire dataset concurrently. Due to this, batch learning is suitable here because unlike online learning, the dataset does not continues to receive new data in real time during the learning phase. In batch learning, the model is trained and evaluated on the entire amount of data which results into better performance as stated by Vulpescu and Mihai Beldean in 2024.

**Critical Considerations:**  
To this end, the project is based on four major concepts: determining learning objectives; identifying the kind of learning process; knowing the task type; and understanding the actual objective. These considerations will help towards making predictions using the model both precise, and explainable. Furthermore, since the leanings concentrate on the practical recommendations, the discoveries can quickly be utilised to influence business decisions, which adds the functional incorporation significance to the project, making for an academically as well as commercially viable project.

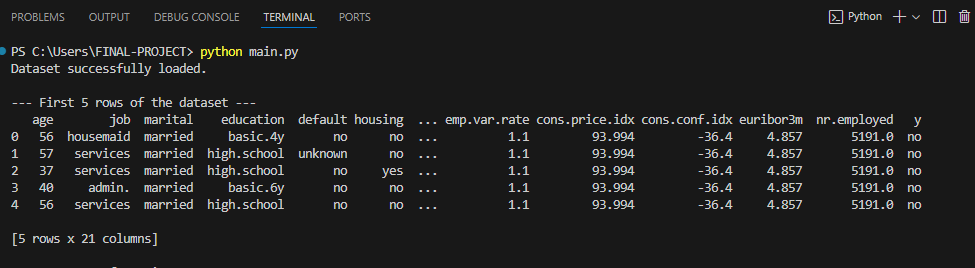
**Data Loading**

The first operation in this project is to import the dataset into a Python environment in which we will analyze it. The data set is available in the file bank-additional-full.csv data format which is semi-colon delimited-CVS. To load the data, we will make use of a Python algorithm lab known as Pandas that is great for data manipulation and analysis. The pd.read\_csv() function will be used to at this data, the delimiter will be set to a semi-colon (;) to possibly capture all the data appropriately. It will also be an active process including, for example, checking for possible problems with data accessibility and correct formatting at the loading phase. Several checks will be made to confirm that the dataset was loaded correctly several checks will be conducted. For instance, when using this function to analyze the data, we will use the head() function to tweak a small section of this data to be sure that it’s processed correctly.

Code snippet implementation

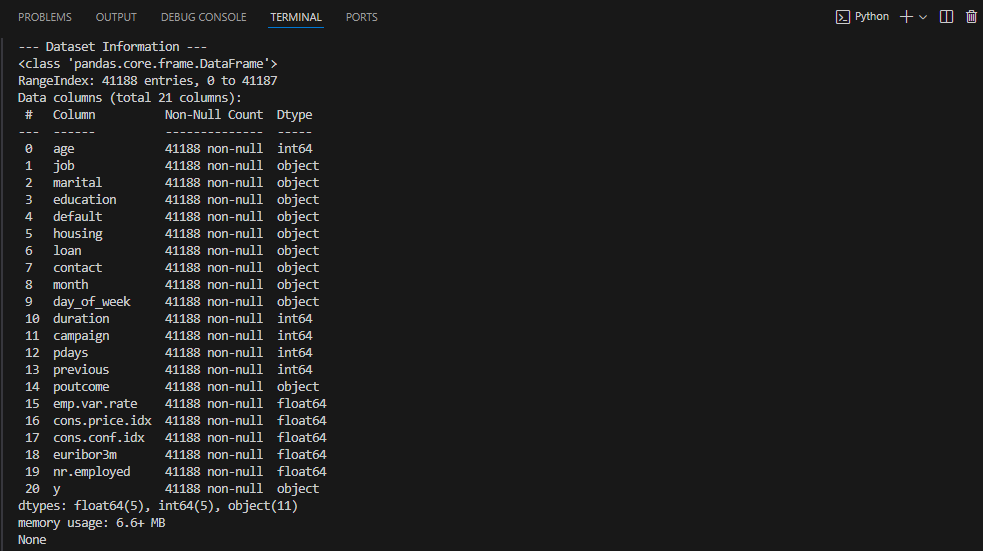


Dataset successfully loaded using **read\_csv()**



Additionally, the info() function will provide a summary of the dataset, including the column names, data types, and the presence of null values. This step ensures the dataset is ready for further analysis.

Summary of the dataset using **info()** function



**Assumptions About the Data**

The following assumptions are made for this analysis regarding the structure and variable characteristics of the dataset. The features in the dataset are a mixture of categorical and numerical features Thus we can start by summarizing the data. For example, age, duration, and campaign are numerical fields and job, marital and education are categorical fields. This means that there is an expectation that whatever is categorical is stored as string, and numerics are stored as such in the dataset. There are special value “unknown” in several columns of dataset where missing data is represented. These will be considered as missing values while data is being preprocessed. Furthermore, as in previous approaches, we assume that the output variable y is binary and correctly labeled “yes” for clients who subscribed to a term deposit and “no” for those who did not.

**Initial Data Exploration**

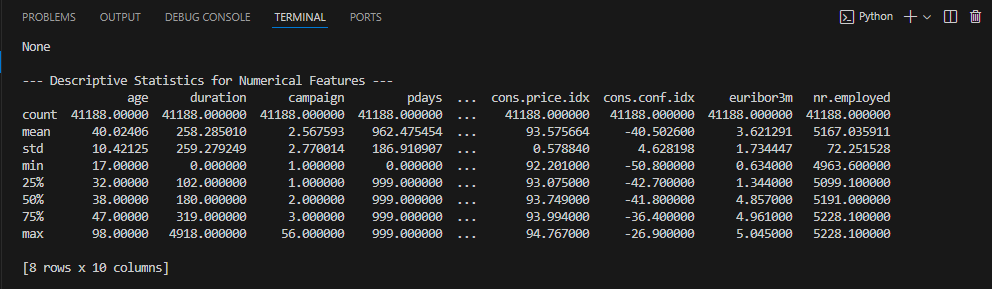
Based on research by (Mishra et al., 2023), exploratory data analysis should be done prior to data analysis and modeling. This involves a look at its physical structure, at the information it contains as well as some of the most fundamental measures of its properties. The first step will be to know the layout of the dataset that will be achieved by invoking the shape attribute that will display the size of the rows and columns. This assists to set the size of the data and confirm each record is in the data set.

Dataset Shape using **shape** attribute



We will also use the info() method to check the data type of each column as well as check for missing values INFO(). It is important to understand the data types before preprocessing because the treatment applied to numerical and categorical data is different. The describe() function will then be used to generate descriptive statistics of the numerical column means median, minimum, maximum which will act as an indication of range of these attributes.

Descriptive Statistics using **describe()**



A visual check is just as significant when trying to identify any patterns in the data. In the case of head() function we will be able to look into the first rows of the distributed dataset and check its format and structure according to our expectations. Also, tail() function will be applied to check the last few records in order not to miss some errors such as tails of null values or additional records. Together these steps provide assurance of the quality of the data at hand, and place the data in a condition to be used in the next analysis stages.

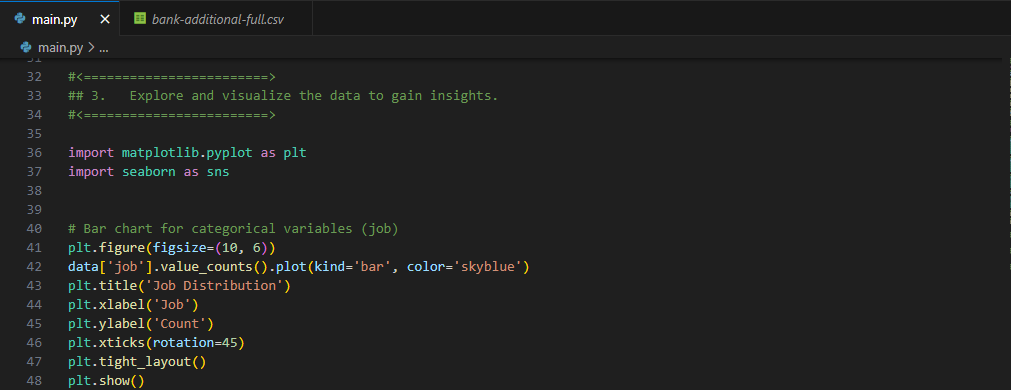
**Data Exploration and Visualization**

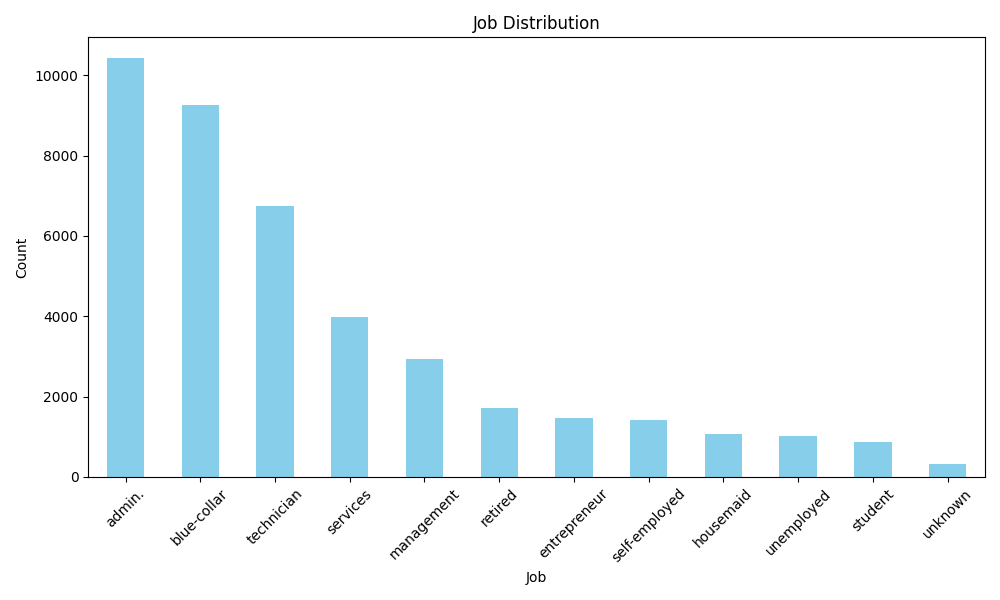
Data exploration and data visualization are reasonable techniques of data investigation, data profiling, and discovering any anomalies that can impact data modeling (Chatzimparmpa et al., 2020). Coherent picture is drawn when we do visualization and statistical analysis of the data which helps to understand impact of variables on each other, the important patterns to be observed and finally to make effective decisions about preprocessing the data as a result of exploratory data analysis. This part of the paper emphasizes the graphical and statistical analysis of the results found in the dataset.

**Graph Generation and Insights**

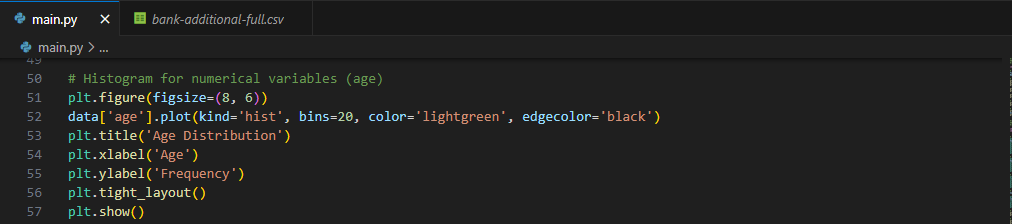
To explore and visualize the data, we will generate four types of graphs: a bar chart, a histogram, a box and whisker plot, and a heat plot. That is, these visualizations enable researchers to focus on various aspects of the data to provide insights.

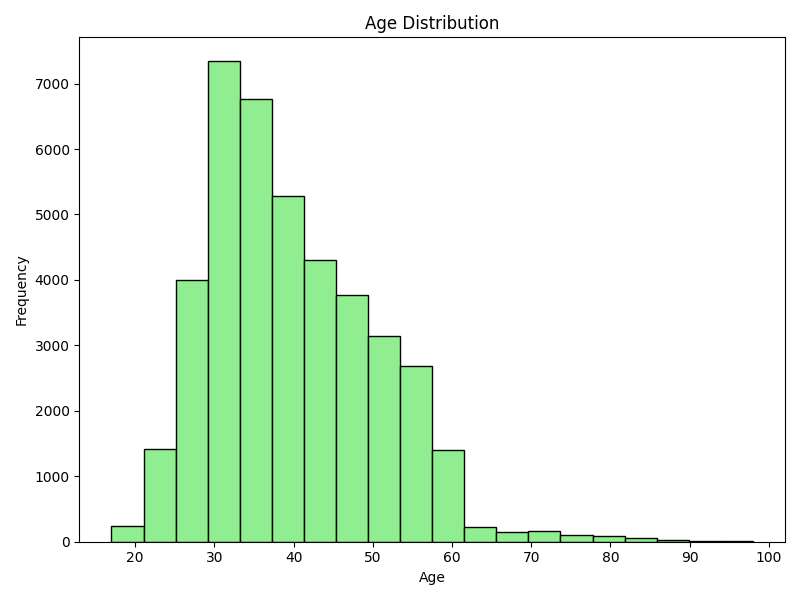
**Bar Charts for Categorical Variables:**  
Bar charts will be used in representing relative frequencies of the categorical variables like job, education and marital status. These graphs can be used to be aware of most well-represented categories and to assess whether there are any biases in the data. For instance, distribution of the customer’s job or education level shows, which segment of the population is utilizing the services of the bank.



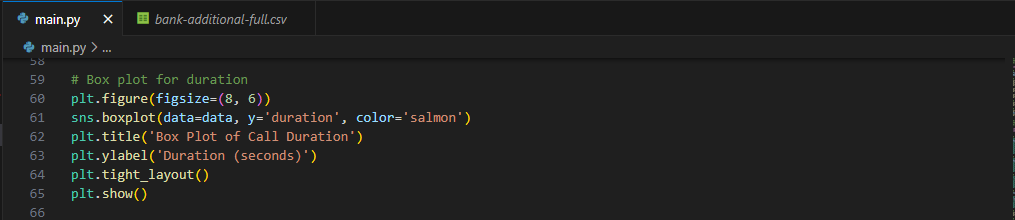


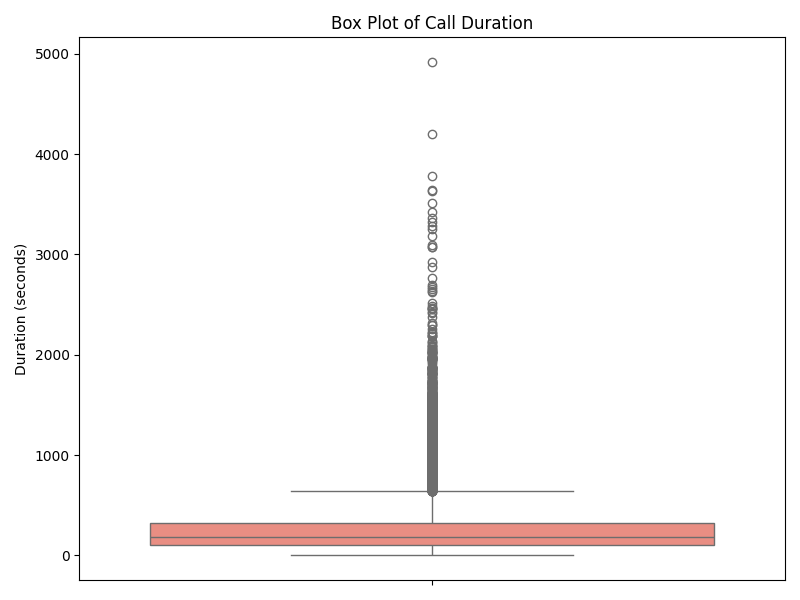
**Histograms for Numerical Variables:**  
Histograms are used in presenting the general appearance of numerical variables such as age, duration and campaign. For example, the age histogram will show the distribution of customers by age and thus whether certain ages prevail in the dataset and if they have certain behaviors.





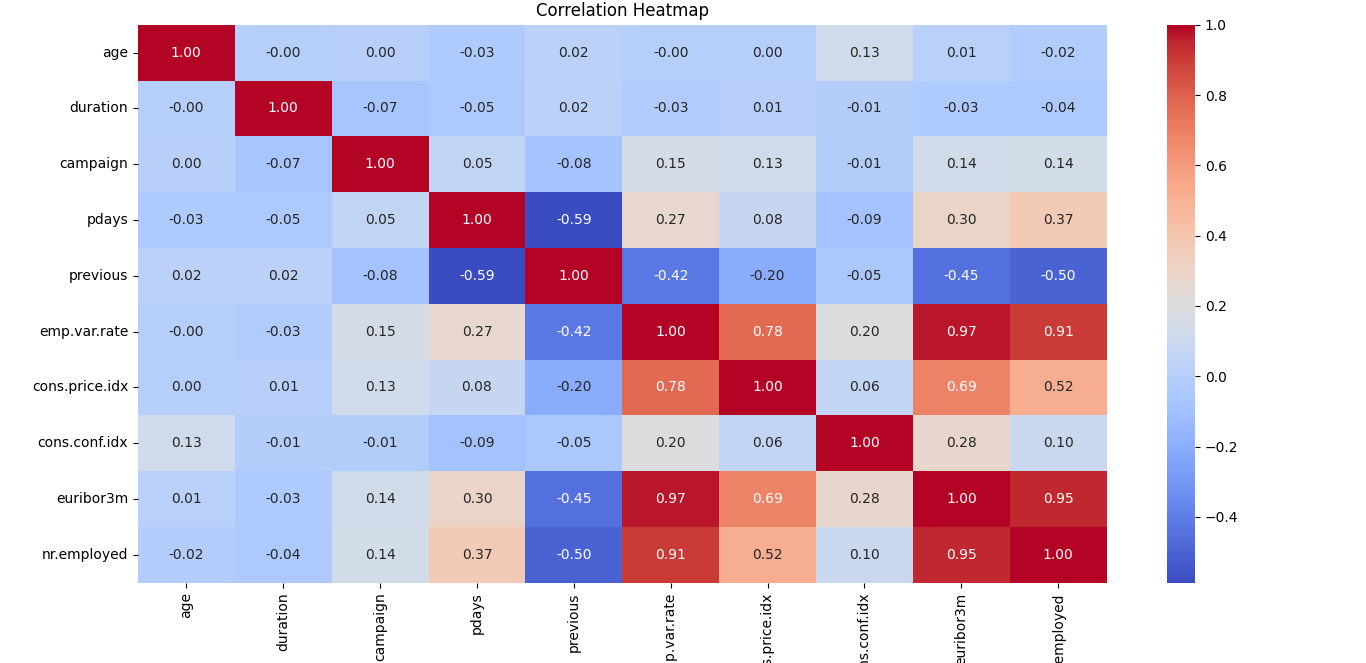
**Box Plots to Detect Outliers:**  
Outliers are very important in case of data that are used to build models, and thus, the box plots for variables such as duration and campaign will be very useful. For instance abnormally high values in duration may suggest that calls are taking long and are thus are affecting the output variable (y) more than warranted.





**Correlation Heatmap:**  
Therefore, the correlation matrix of numerical variables: age, the duration, and euribor3m will help in determining how strong these relationships are. Diagnostic measures can be used to decide on preprocessing steps and detect multicollinearity, a helpful sign in feature selection.



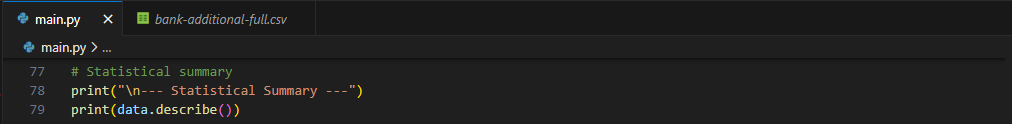


**Highlighting Key Findings**

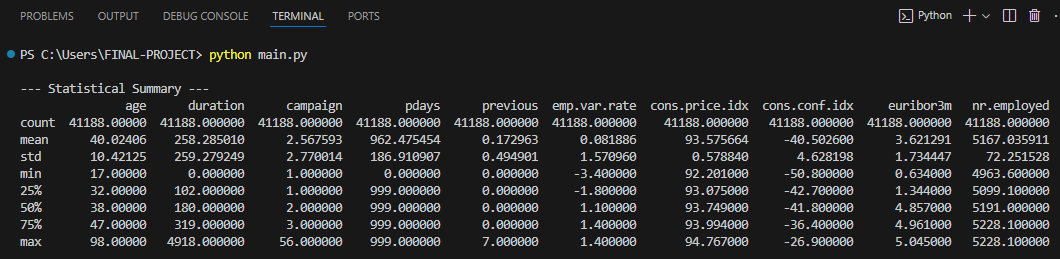
From the exploratory visualizations, we aim to highlight the following findings:

1. **Imbalance in the Target Variable (y):**  
   This bar chart of y will show whether there is a sloping of the graph slope between the “yes” count and the “no” count. There may also be some inequality during modeling because resampling techniques may be required in case of a huge disparity.
2. **Call Duration's Impact on Subscription:**  
   Subsequently, customer response analysis with respect to duration will reveal the kind of pattern expected such as longer call duration implies high subscription rate.
3. **Distribution of Demographics:**  
   Frequency distributions of age and occupation and grouped frequency distributions for job and education level will give information about the characteristics of the customers, including details which could be used in targeting.
4. **Economic Indicators and Subscription Likelihood:**  
   Correlation heatmap would assist to analyze correlation with the help of such values as emp.var.rate, euribor3m, nr.employed, and the target variable. They can point to the fact that outside stimuli may play a huge role in customer decision making.

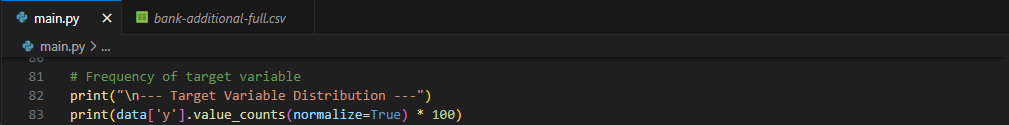
**Statistical Analysis for Deeper Insights**

Statistical analysis helps extend visualization, as numbers allow to consider relationships and variations in the data more comprehensively compared to a visualization alone (Gewers et al., 2022). Mean, median, standard deviation and correlation coefficients etc will be used to analyze numerical variables. These give information on means, spread and association for selection of variables for modeling Midstream: In midstream, more in-depth description and quantification of variables is carried out (Gewers et al., 2022).

Statistical summary using **describe()**



For categorical variables, given variables will be tested for their mode and modal groups through construction of appropriate tables: Frequency tables. The relationships between the categorical predictor variables and y can be viewed from the cross-tabulations between categorical variables and the target variable. For example, Job will show which professions are more connected to y.



Target Variable Distribution using **value\_counts()** function



**Handling Outliers and Anomalies**

Skews can also pose a problem when the data contains such unusual features such as outliers that can disrupt modeling outcomes. Capping or removal techniques would be applied to outliers, which can be defined with the help of box plots (Ali et al., 2023). For example, high values of the duration or campaign can be limited to maximum value specified by the interquartile range.

**Feature Correlation and Patterns**

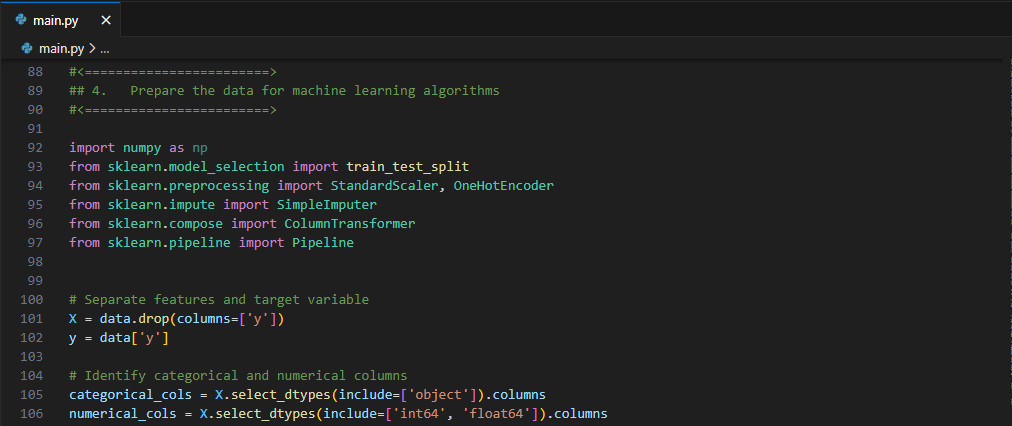
Hossain et al., (2024) rightly pointed out that the correlation heatmap is used to identify variables that are highly associated with the target variable or one another. This way features with high multicollinearity level should be either removed or transformed for better performance of SVM model. For instance, it might be possible to delete one of the two variables, emp.var.rate and nr.employed if they are highly correlated. Patterns including seasonality in month or day\_of\_week and others like Demographic/Economy using visualizations and statistical methods. These insights pave way for the design of the preprocessing techniques such as feature choice as well as encoding techniques.

**Data Preparation for Machine Learning Algorithms**

The paper by Janiesch et al. (2021) posits that, effectiveness of machine learning models highly depends on the input data information quality and data characteristics. Several fundamental steps must be followed when preparing data for machine learning: pre-processing, cleaning, making decisions about outliers, data partitioning and checking robust preprocessing across data sets. Each step makes the data more clean and consistent for the algorithms to develop promising and reliable prognosis.

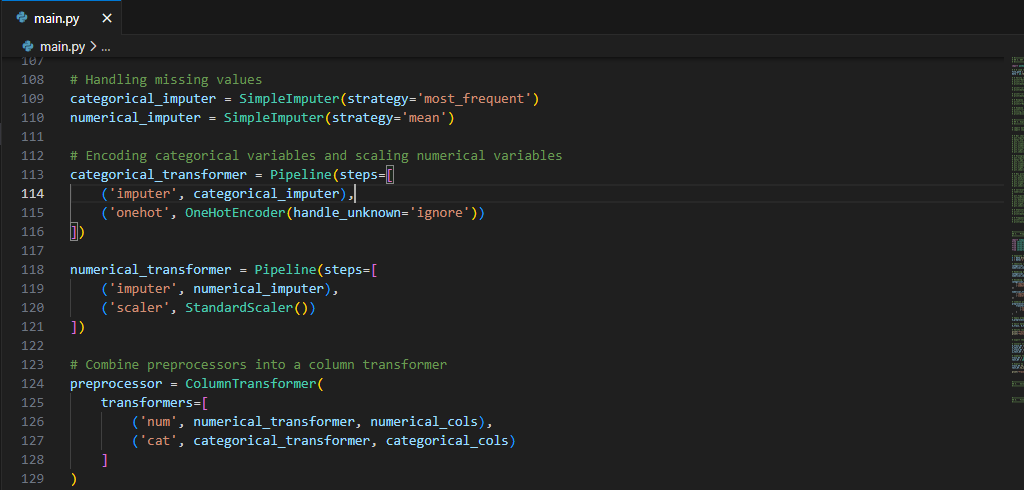
**Pre-Processing the Data**

Pre-processing prepares raw data for analysis by the machine learning technique to be used in the analysis. This involves fixing features that have messed up either through inconsistencies, encoding categorical variables, scaling numerical variables, and standardizing features (Cristina, 2021). The data pre-processing technique does not only serve the purpose of data quality improvement but also the purpose of Model Quality improvement and Interpretability. Encoding is ever important when working with categorical-binary variables, in order to transform text labels into numerical forms. Tasks such as one-hot encoding make it possible for machine learning algorithms to process categorical data without making assumptions about a sequence by default (Fernando & Morales, n.d.). For instance, the feature such as job might look like job1, job2, job3 and so on where every job type is described by a different binary column. Attribute normalization is done to bring the numerical features into the same range and to solve the problem of certain features being dominant because of large values. Some of the widely used techniques include standardization which places data having a mean of zero and a standard deviation of one, and normalization which places data in the scale of 0 to 1 respectively. Feature selection is also a pre-requisite of pre-processing since feature plays an important role in determination of the success of a model (Zebari, Bahsoon, & Alomary, 2020). Correlations and variance detect these features as irrelevant or redundant and removes them in order to eliminate noise in big data and to quicken computations.



**Data Cleaning**

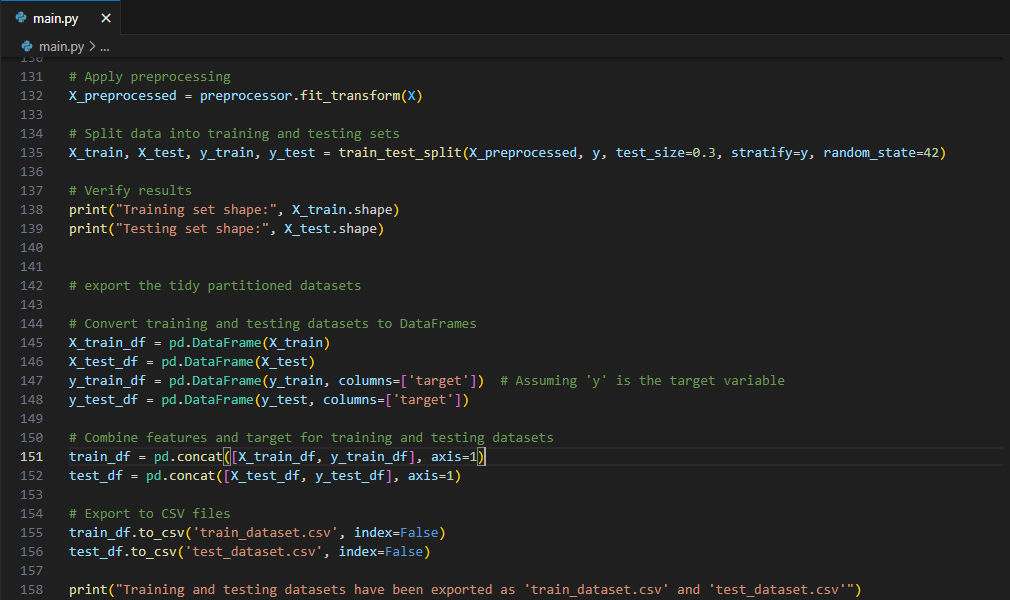
The cleaning process ensures that problems, which are inherent in the dataset such as duplicate, inconsistent, and missing data are detected and addressed (Cho et al., 2020). If not well managed, these problems affect model performance in a considerable way. For instance, they distort the actual weight of the information hence making the predictions biased. Data missing is one of the biggest threats in machine learning (Whang et al., 2023). The discrepancy can be caused by a mistake in the data acquisition process, or by lack of comprehensive data. To manage this, the different methods we use include imputation or deletion. This is particularly because imputation techniques such as using mean, median or mode ensures that data is not lost. There are more innovative techniques, for instance, predictive imputation by means of regression, in which other features compute predicted values of missing data (Cho et al., 2020). Typically working with categorical variables, cleaning implies dealing with issues of either/sprinkling like different spelling or capitalization. For example values such as “Manager”, “manager”, “MANAGER” and so on are sorted in a standard format.



**Handling Outliers**

An example that Wada (2020) has mentioned is that outliers can have an impact on the statistical summaries and model efficiency. For example, a large value in the frequency feature could polarise the training of the model. Outliers identification includes graphs which are the box plots and the statistical measurement known as interquartile range (IQR). So as soon as they are identified it can be taken care of through capping or removal. In capping, the we assigns equivalent highs and lows and removes values outside the expected range in order to normalize them. The removal approach is less complicated, but may lead to discarding potentially useful data. In this case, depending on whether the data is important and how crucial the feature is in this set, the determination will be made. Handling Outliers aims at minimizing the impact of extreme data points in a dataset such that the final modeling increases its generalizability.

**Data Partitioning**

Partitioning of data into the training and testing sets is considered to be the initial process that has to be done in the machine learning. It also helps in a blind assessment of how well models would generalize and tends to overcome issues like, overfitting whereby a model provides very favorable results on training data but very poor results on fresh data (Kernbach & Staartjes, 2022). Normally the dataset should be divided with 70% training data and 30% testing data. Nevertheless, ratios of coefficients of consecutive layers can change depending on dataset size. For instance, big datasets may split the data 80/20 training testing split so as to capture as many training data points as possible while having adequate testing data (Kernbach & Staartjes, 2022). Organizing the target variable allows stratified sampling to retain the distribution’s predictability and coherence in the two subsets during class-imbalanced scenarios (Kernbach & Staartjes, 2022).

**Ensuring Robust Data Preprocessing**

Like in all the data preprocessing methods, guaranteeing a robust process is important in feature selection. Feature scaling is a significant practice especially with SVM or kNN models since distances are calculated (Dremel et al., 2022). One complaint of scaling is that, for example, larger values of features will overpower the small values of other features and the model prediction will be influenced as a result. To resolve this problem normalization and standardization are used.

As a standard procedure, the same preprocessing transformations that were applied on the training set must also be applied on testing set (Bansal et al., 2022). These methods are often done automatically through StandardScaler in Python’s sklearn library where in training, necessary parameters like mean or standard deviation are stored and used in the test script.

There is always a compromise when one has to deal with the operation of missing data. However, deleting rows with missing values substantiates a neat dataset foundation while at the potential of lowering it through expunging samples with high missing value percentages. On the other hand, imputation retain the structure and size of a dataset: the differing values are kept also with their errors and possible biases are introduced. A relevant importance of evaluating these methods is to determine the influence of these methods towards the model performance so as to select the right technique.

The use of preprocessing is established using the iterative model training and the model testing process. These factors consist of accuracy, precision, recall, and F1-score all of which define the effect of preprocessing on model indicators. Preprocessing strategies are tested one more time in cross-validation, which checks the stability of these steps and the results’ similarity for different data sections (Saidani et al., 2024).

**Selecting and Training the Machine Learning Model**

The process of developing a machine learning model is selection, training and then the model is tested. This process included familiarisation of the data set, defining the purpose of the exercise, and identifying the best model for the problem given the strengths of the selected model. Furthermore, the interactions such as the problems of overfitting, underfitting, as well as the questions of hyperparameters settings cannot be neglected to guarantee reliable results. This is explained here together with the code and the following are the steps:

**Model Selection**

Selecting a model is a very important process in the ML pipeline. This will depend on whether the problem is a classification problem or a regression problem, the structure of the data set and the objectives, of the project. For this predictive modeling task, a classification model is required, as the target variable (y) represents a binary outcome: whether a particular client will take up a term deposit. Owing to the characteristic of the used dataset and the task to be solved, logistic regression was selected as the baseline method. Logistic regression is easy to explain, less time consuming and was proven to be effective in classifying data into two classes as perceived from the experiment. Other comparably used models were decision trees, support vector machine (SVM), as well as ensembling methods such as Random forest or Gradient boosting machines (GBM). Though these models are more elaborate and could be potentially more effective, decision to start with logistic regression was made primarily in order to set the bar due to its relative simplicity and genetic predisposition to the phenomenon known as overfitting.

**Performance Measure**

The performance of the model is then assessed based on metrics which are relevant towards objectives set in this project. Hence, the following accuracy, precision, recall and F1 score are adopted for this binary classification problem.

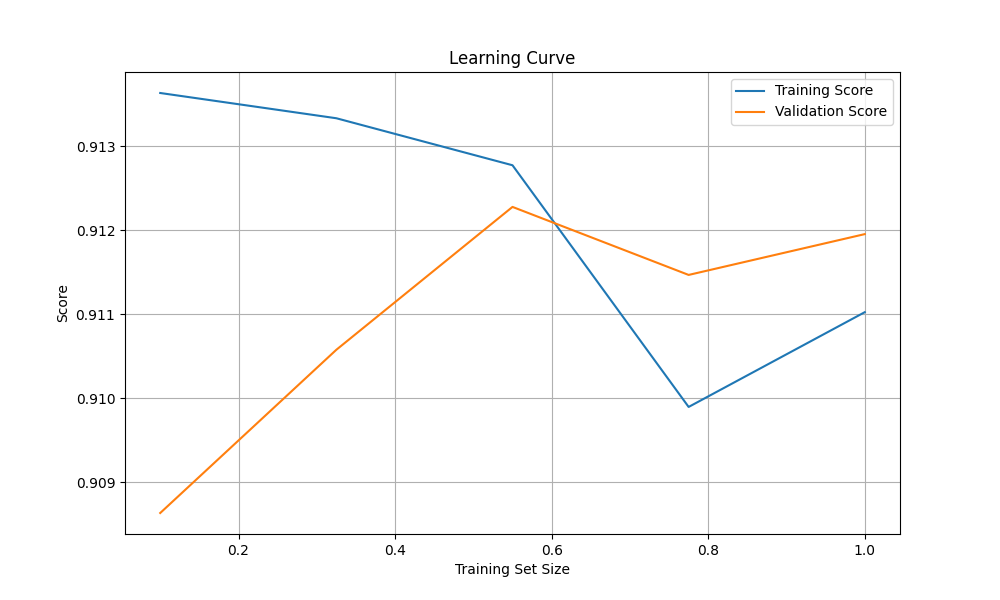
1. Accuracy, thus, gives an approximation on how right the results are.

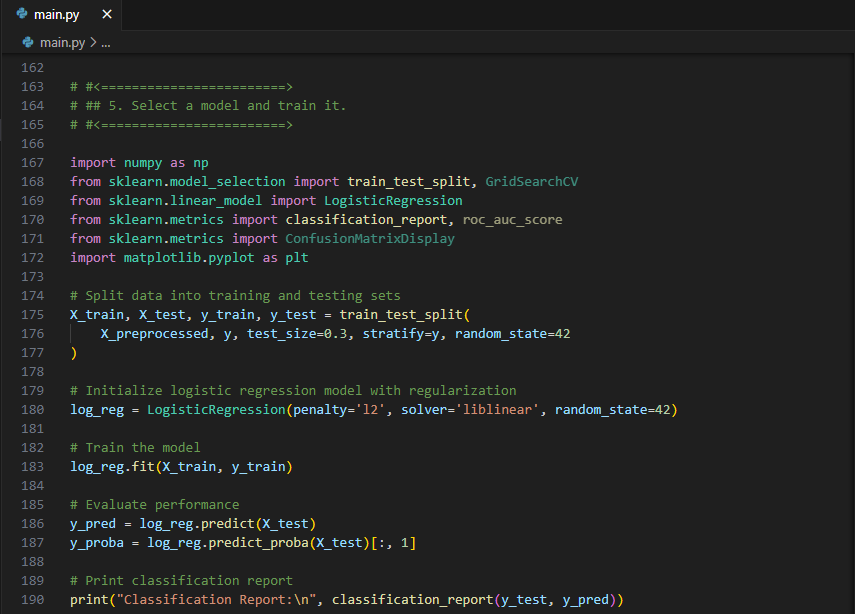
2. Precision and Recall are concerned with how well the model correctly identifies the positive class, which is indispensable when considering the results of the marketing campaign.

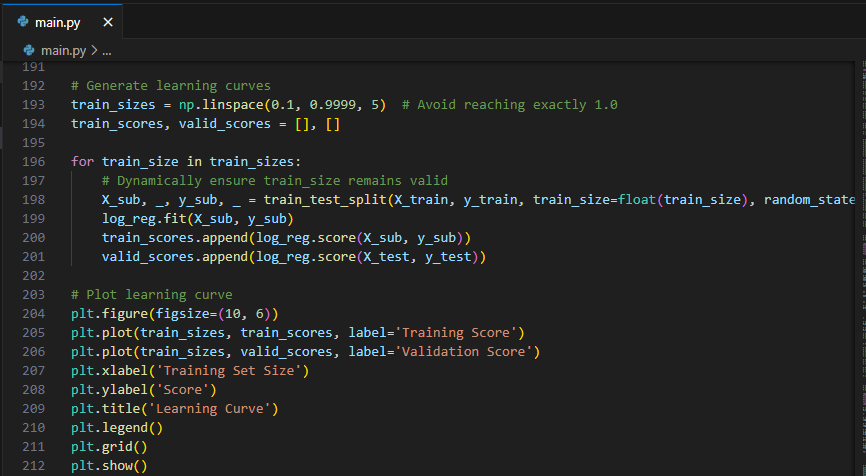
3. Precision being expressed as Tp / (Tp + Fp) and Recall expressed as Tp / (Tp + Fn), F1 score which is the average of the two is used when the dataset is imbalanced since the coefficient is in between 1/Sensitivity + 1 and 1/Specificity + 1. The selection of the metrics guarantees the coverage of the model’s positives and negatives. For instance, while precision defines false positive results, recall points out false negatives, critical aspects of evaluating the success of a campaign.

**Learning Curve Graph**

Concepts of learning curve involves demonstration of the training and validation errors with reference to the amount of data through a number of cycles. They are useful when it comes to diagnosing between overfitting and underfitting (Montesinos López et al., 2022). Ideally, the training error should be low and the validation error should be high in order to get a better model possibly oversensitive. On the other hand if both of them are high then the model can be over fitting. From the learning curve, modifications in the model in terms of the model’s difficulty or a larger dataset can be done to fine tune it (Montesinos López et al., 2022). For this task, learning curves are commitment by training the logistic regression model using different portion of the dataset and measure the performance based on the training set and the validation dataset.







**Overfitting and Underfitting**

In feature selection, overfitting is a major problem where the model learns the noise and idiosyncrasies of the training set rather than generating a robust hypothesis set. It can be identified by a large difference between the minimal training error and the relatively high validation error. This group includes methods that avoid high penalty for large coefficients, for instance L1 or L2 penalties reduce the degrees of freedom, hence preventing over fitting (NWOSU et al., 2024). When the selected model cannot capture the right patterns in the data then it a situation referred to as underfitting whereby the model will return high errors both for training and validation data sets. For the logistic regression regularization works by using C, which stands for the regularization parameter in sklearn. As C decreases in general, branching of regularization is supported, which decreases the likelihood of overfitting.

**Regularization Techniques**

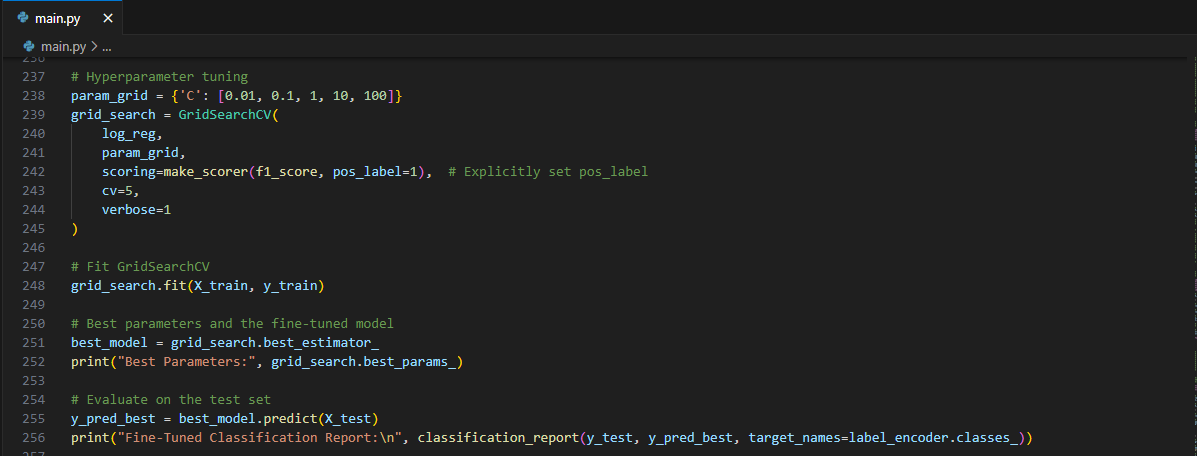
The work of Regularization is to minimize the complexity of learnt model so as to enhance the generalization capability. Here in this task for avoiding large value of coefficient Ridge regularization (L2) is used. This technique helps restrict the size of the model and hence enhance model’s ability to perform well for data it has not encountered before. Such process is very essential especially when we are dealing with large data or data with features that share high correlations.

**Hyperparameter Optimization**

In fact, hyperparameters are used to regulate model capacity and learning process parameters. For logistic regression, the only hyperparameter is the regularization strength or concentration (C). recursive feature elimination is used while the value of C is tuned using grid search and cross-validation. This helps the model to have betterBias/Variance trade-off thus giving better generalization. Tuning of hyperparameters for ensemble models such as Random Forests, consists of parameters including the number of decisions trees, the maximum tree depth, the minimum number of instances per split. The machine learning models are used with cross-validated grid search or randomized search techniques in order to select the best configuration.

**Fine-Tuning the Machine Learning Model**.

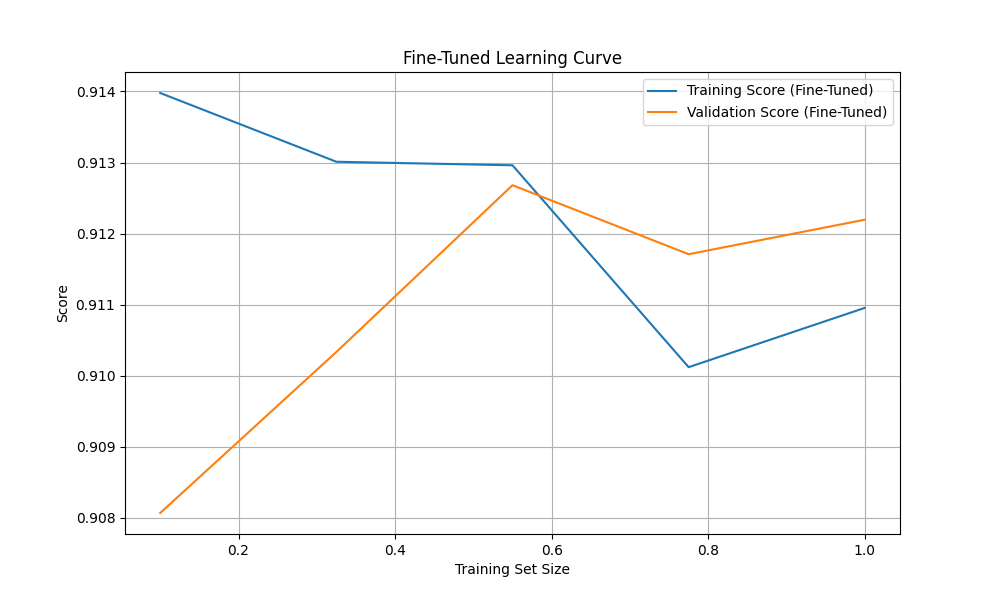
**Hyperparameter Tuning**

Due to the factor which is hyperparameters tuning is a very crucial step for improving the performance of a model. For logistic regression, the only adjustability hyperparameter is the strength of regularization (C) which regulate the quantity of training error and the model complexity directly. Operation of tuning is done by repeating the process of the grid search over different C values aiming at balancing bias and variance. Hyperparameter tuning is conducted with the use of cross validation so as to select the hyperparameters most suitable for generalization. Optimization of more complex models for example: Random Forest or Gradient Boosting would include parameters as number of estimator, maximum depth of the trees and learning rate. Compared to the above hyperparameters, these dramatically affect the models’ complexity in identifying fine subtleties of the data.

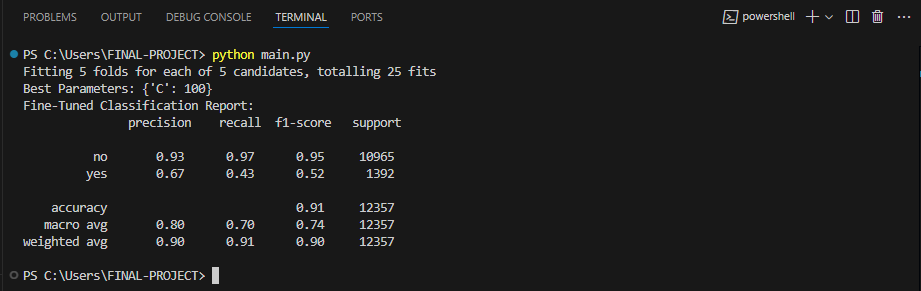
**Regularization Adjustments**

In this phase, regularization, which was used in the preliminary model training phase with the L2 (Ridge) technique, is further refined. The C parameter in the problem affects the generalization of the decision boundaries by helping preventing underfitting and overfitting. The way of the combination of the regularization approaches lets reveal significant patterns rather than overfitting the model and overdetermination to training data. This step is notably crucial in the data with numerous features where the collinear features exaggerate the coefficient values.

**Evaluation Metrics for Fine-Tuning**

Some of the performance evaluation metrics for fine-tuning includes speaks for better understanding. During fine-tuning, performance is measured by accuracy, and precision, and recall and F1-score are used. Such statistics are accompanied by the area under the receiver operating characteristic curve (AUC-ROC) that gives an overall idea of the model performances at different classification margin. Progress made on these aspects during fine-tuning suggests that the model’s performance is more in tune with the goals of the project.

Interpretation of the metrics involve the comparison of epoch results between the training and validation set, to identify neural networks that have overfitted its training set or underfit the entire set. Only in this case generalization is achieved – a significant increase in the validation performance while the training performance is not increased proportionally.



**Recommendations**

The problem was addressed in the model due to the high accuracy achieved and the potential to reduce the incidence of poor customer response and to optimize resources such as time and money utilised in the marketing campaigns by banks. Some suggestions for deployment include the need to carry out the model in a production environment and follow the trends in real-time applications. Moreover, future work may also try to extend the ensemble approach to improve the model’s stability and flexibility. It was also noted that with enhanced feature engineering and, quite possibly, better handling of the problem of minor class imbalance, the accuracy could be brought to marginally higher levels. In general, I believe that making use of a structure combined with constants cyclical enhancements were instrumental in achieving the objectives of the project.

**References**

Ali, M., Zhu, P., Ma, H., Pan, H., Abbas, K., Ashraf, U., Ullah, J., Ren, J., & Zhang, H. (2023). A Novel Machine Learning Approach for Detecting Outliers, Rebuilding Well Logs, and Enhancing Reservoir Characterization. *Natural Resources Research*, *32*(3), 1047–1066. https://doi.org/10.1007/s11053-023-10184-6

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, *3*(100071), 100071. https://doi.org/10.1016/j.dajour.2022.100071

Chatzimparmpas, A., Martins, R. M., Jusufi, I., Kucher, K., Rossi, F., & Kerren, A. (2020). The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum*, *39*(3), 713–756. https://doi.org/10.1111/cgf.14034

Cho, B., Dayrit, T., Gao, Y., Wang, Z., Hong, T., Sim, A., & Wu, K. (2020). Effective Missing Value Imputation Methods for Building Monitoring Data. *2020 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata50022.2020.9378230

Cristina, S. (2021, July 7). *The Role of Applied Behavior Analysis in Increasing Employees’ motivation and Productivity*. https://www.researchgate.net/profile/Salagean-Cristina/publication/359867254\_The\_Role\_of\_Applied\_Behavior\_Analysis\_in\_Increasing\_Employees

Fernando, Á., & Morales, K. (n.d.). *On the encoding of categorical variables for machine learning applications TESIS QUE PARA OPTAR POR EL GRADO DE MAESTRO EN CIENCIAS E INGENIERÍA DE LA COMPUTACIÓN PRESENTA: ERICK GARCÍA RAMÍREZ*. https://ru.dgb.unam.mx/bitstream/20.500.14330/TES01000806508/3/0806508.pdf

Gewers, F. L., Ferreira, G. R., Arruda, H. F. D., Silva, F. N., Comin, C. H., Amancio, D. R., & Costa, L. D. F. (2022). Principal Component Analysis. *ACM Computing Surveys*, *54*(4), 1–34. https://doi.org/10.1145/3447755

Hossain, N., Anjum, N., Alam, M., Rahman, M. H., Md Siam Taluckder, Al, V., Islam, S., & Afrin Hoque Jui. (2024). PERFORMANCE OF MACHINE LEARNING ALGORITHMS FOR LUNG CANCER PREDICTION: A COMPARATIVE STUDY. *International Journal of Medical Science and Public Health Research*, *05*(11), 41–55. https://doi.org/10.37547/ijmsphr/volume05issue11-05

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(31), 685–695. Springer. https://doi.org/10.1007/s12525-021-00475-2

Kernbach, J. M., & Staartjes, V. E. (2022). Foundations of Machine Learning-Based Clinical Prediction Modeling: Part II—Generalization and Overfitting. *Acta Neurochirurgica*, 15–21. https://doi.org/10.1007/978-3-030-85292-4\_3

Mishra, A., Soni, U., Arunkumar, A., Huang, J., Kwon, B. C., & Bryan, C. (2023, April 4). *PromptAid: Prompt Exploration, Perturbation, Testing and Iteration using Visual Analytics for Large Language Models*. ArXiv.org. https://arxiv.org/abs/2304.01964

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, Model Tuning, and Evaluation of Prediction Performance. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, 109–139. https://doi.org/10.1007/978-3-030-89010-0\_4

NWOSU, A., Imuetinyan, G., Ajayi, B. A., & Morufu Olalere. (2024). The Impact of Regularization on Linear Regression Based Model. *Journal of Artificial Intelligence and Computer Science*, *1*(1). https://doi.org/10.5281/zenodo.13118703

Rony, M. A. T., Hassan, Md. M., Ahmed, E., Karim, A., Azam, S., & Reza, D. S. A. A. (2021). Identifying Long-Term Deposit Customers: A Machine Learning Approach. *2021 2nd International Informatics and Software Engineering Conference (IISEC)*. https://doi.org/10.1109/iisec54230.2021.9672452

Sağlam, M., & Montaser, S. E. (2021). The Effect of Customer Relationship Marketing in Customer Retention and Customer Acquisition. *International Journal of Commerce and Finance*, *7*(1), 191–201. http://ijcf.ticaret.edu.tr/index.php/ijcf/article/view/259

Saidani, O., Umer, M., Alturki, N., Alshardan, A., Kiran, M., Alsubai, S., Kim, T.-H., & Ashraf, I. (2024). White blood cells classification using multi-fold pre-processing and optimized CNN model. *Scientific Reports*, *14*(1), 3570. https://doi.org/10.1038/s41598-024-52880-0

Vulpescu, E., & Mihai Beldean. (2024). Optimized Fine-Tuning of Large Language Model for Better Topic Categorization with Limited Data. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-4697966/v1

Wada, K. (2020). Outliers in official statistics. *Japanese Journal of Statistics and Data Science*, *3*. https://doi.org/10.1007/s42081-020-00091-y

Whang, S. E., Roh, Y., Song, H., & Lee, J.-G. (2023). Data collection and quality challenges in deep learning: a data-centric AI perspective. *The VLDB Journal*, *32*(4). https://doi.org/10.1007/s00778-022-00775-9

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56–70. https://doi.org/10.38094/jastt1224