Portfolio Project

Pinaki Pani
Student ID: 23112573
Data Mining and Machine learning 1
Master of Science in Data Analytics, National College of Ireland
x23112573@student.ncirl.ie

***Abstract*—This project encompasses three distinct but interconnected projects, each focused on leveraging machine learning techniques to address critical challenges in different domains. In the first part the surge in global energy consumption over recent decades is analysed which underscores the significance of accurate forecasting in managing and conserving energy in smart buildings. This study introduces a novel machine learning (ML) methodology for predicting energy usage in industrial structures. The second part delves into the Adult Census dataset, a comprehensive collection of demographic and socio-economic attributes. The primary objective is to build a robust machine learning model capable of accurately classifying individuals' income levels based on certain attributes.The third part revolves around the Bank Marketing Dataset, focusing on predicting whether a client will subscribe to a term deposit in a Portuguese banking institution. The outcome is expected to provide valuable insights, improving the success rate of term deposit subscriptions. Collectively, these projects showcase the versatility and efficacy of machine learning in addressing complex challenges across industries.**

PART A: Steel Industry Electricity Consumption Dataset

I. Introduction

Energy is recognized as a crucial factor for economic and social development. In recent decades, the energy challenge has intensified, primarily due to advancements in machinery relying on electricity. Consequently, accurately predicting the energy consumption of buildings becomes a pivotal aspect of energy conservation in the field. The prognosis of machine tool energy usage plays an indispensable role in planning, managing, and conserving energy within the manufacturing industry [1]. Accurately categorizing the power consumption with the help of other available information regarding the electricity consumption has become immensely significant in the current electronic world, especially for industries that considers each and every factors as it plays a crucial role in capital management of their business.

A. Objective

The objective of this project is to develop and deploy machine learning models for predicting and optimizing energy consumption in the steel industry. Leveraging statistical analysis, exploratory data analysis (EDA), and advanced machine learning techniques the project aims to achieve the certain goals. First to learn and develop accurate models to predict energy consumption in the steel industry based on various input parameters. Next to compare the performance of different ML models to identify the most efficient and effective approach for predicting energy consumption in the steel manufacturing process.

The project aims to contribute to more informed decision-making in the steel industry, ultimately leading to improved energy efficiency, reduced environmental impact, and enhanced sustainability.

II. Related Work

Early research on Decision Trees, notably Breiman's seminal work on Classification and Regression Trees (CART) [Breiman, 1984], laid the foundation for employing decision-based algorithms in predictive modeling. Decision Trees have been widely used due to their interpretability and simplicity, making them applicable across various domains. Quinlan's development of the C4.5 algorithm [Quinlan, 1993] addressed challenges associated with Decision Trees, introducing pruning techniques to mitigate overfitting. The evolution to ensemble methods, particularly Random Forests [Breiman, 2001], marked a significant advancement, enhancing predictive accuracy by aggregating multiple Decision Trees. XGBoost, proposed by Chen and Guestrin in 2016 [Chen and Guestrin, 2016], represents a breakthrough in boosting algorithms, showcasing improved scalability and predictive performance. Notably, research by Chen et al. [Chen et al., 2018] demonstrated XGBoost's effectiveness in handling large-scale datasets, making it a compelling choice for this case as well.

III. Data Mining Methodology & EDA

A. Dataset Summary

The dataset totally consists of energy consumption data in the steel industry, since January 1, 2018. It comprises $35,040$ observations and 11 variables. The dataset contains hourly energy consumption readings, capturing variations throughout weekdays and weekends. Variables such as Usage per kWh, Lagging Current Reactive Power per kVarh, and CO2.tCO2. provide insights into energy consumption and environmental impact. Time-related features like NSM, Status of Week, and Day of the week offer temporal context. Upon careful inspection with R's functional attributes it can be written that the all the features except for 'Day of Week', 'Week Status' and 'Load Type' are all numerical columns, where date is of date time format. All the non numerical columns are also categorical type columns as well. With the help of function 'is.numeric' all the

date	Usage	NSM	...	Day_Week	Load_Type
1/1/2018 10:00	3.64	36000	...	Monda	Light_Load
1/1/2018 10:15	4.07	36900	...	Monday	Light_Load
1/1/2018 10:30	3.71	37800	...	Monday	Medium_Load
1/1/2018 10:45	3.60	38700	...	Monday	Light_Load
1/1/2018 11:00	4.21	39600	...	Monday	High_Load

TABLE I: Housing Data

numeric columns are then extracted. The dataset is sampled and shown in Table 1

B. Missing Data

Once the dataset is initially explored and described to check whether the variables are numeric, non-numeric or continuous, next the dataset is checked for the missing values present in it. Upon exploration it is inferred that the dataset does not contain any missing values.

C. Categorical Features Exploration

As inferred earlier that all the non numeric columns are categorical, so now those features are explored and analysed with the help of different r functionality to check if there is any class imbalance. **WeekStatus:**The data within this feature predominantly captures observations on weekdays, with 25,056 instances, compared to 9,984 on weekends. **Day_of_week:**The dataset provides a balanced representation of days of the week, each contributing 4,992 observations. **Load_Type:**The load type distribution reveals three distinct categories: Light_Load represents the majority with 18,072 occurrences. Maximum_Load indicates a significant but lower frequency with 7,272 instances. Medium_Load contributes 9,696 occurrences.

D. Visual Exploration

Within the EDA there are certain visual exploration analysis that provides crucial information about the dataset that can be influential for the data preparation to do modelling over it. The scatter plot is plotted for the numerical features across some other key numerical features that gives effective information for visual analysis. The Leading and Lagging Reactive Power kWh values are plotted across Usage kWh. It is evident that Lagging Power kVarh has a linear realationship with Usage kVarh while Leading Reactive Power kVarh has a non linear relation. Similar non linear relationship can be viewed for the scatter plot for Laggin and Leading Power kVarh and as well as for their power factors as well. The doubled rowed scatter plot with the visual information mentioned is shown in the figure Fig 1.
Moving forward the next comes the countplot for the target variable or also known as dependent variable of this dataset i.e. '**Load_Type**'. The three categories in the feature as mentioned above is now shown with the help of pie chart plot that also shows the counts on top of the respective sections. This shows the visual representation of the class imbalance factor that is being

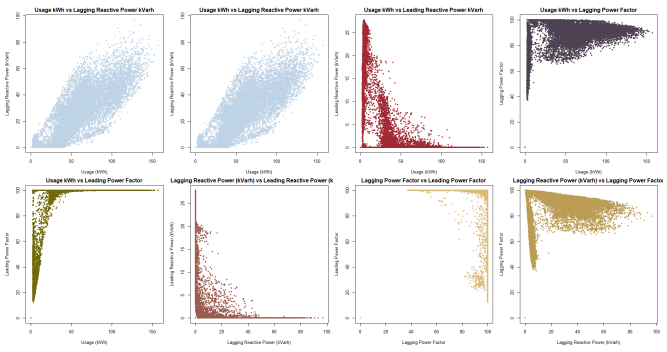

Fig. 1: Scatterplot across Target Variable

analysed here. However there are certainly enough information for each class so that the model can learn enough about their trends. The pie chart is presented in the figure Fig 4. The information for each day of the week is available and is used in the next exploration to check what type of energy consumption occurs during the week days and weekends where the workload and consumption is less compared to weekdays.Upon looking at the countplot for Usage kWh when grouped with the Load Type then the distribution for weekdays and weekends has similar trends but the consumption amount decently differs. The difference in the consumption amount is depicted in the figure for countplot in Fig 2. Among all the visual representation of

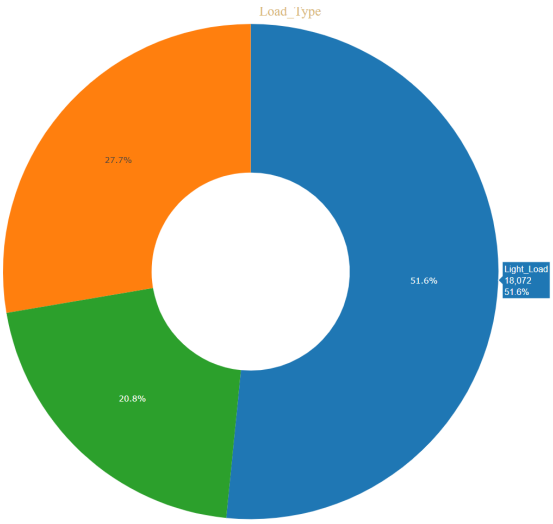

Fig. 2: Pie Chart for Target Variable

interesting facts about the dataset two such inferential plots are scatter plot of the Leading and Laggin Reactive Power consumption vs the Usage per kilo watts, differentiated by the type of load. Colors are assigned to each type of the load and their distribution mixture is plotted for the whole dataset. The Lagging Reactive Power consumption is plotted in the figure Fig 3 and Leading consmption is shown in Fig 4. where it depicts the linaer trend whereas the Leading Reactive Power consumption is plotted in the figure 8 where the non linear trend is apparent.

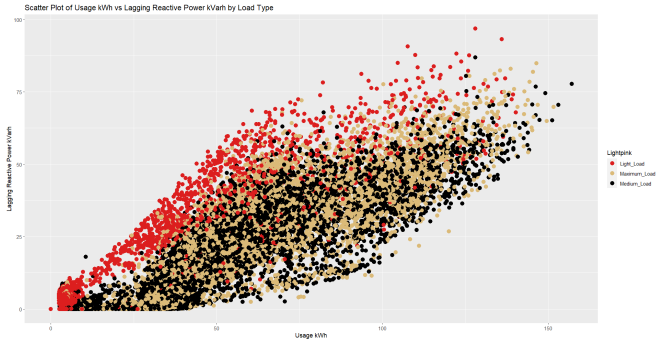

Fig. 3: Scatter Plot For Lagging Power Usage As per Load Type

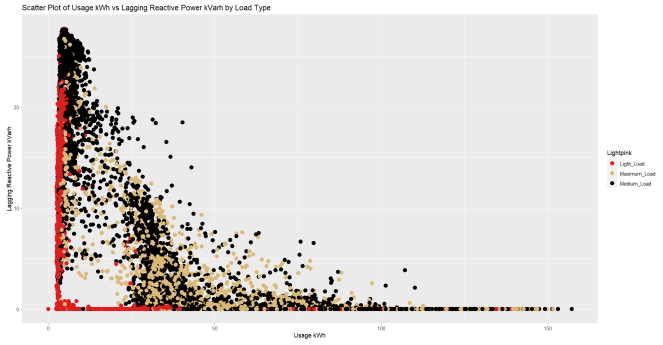

Fig. 4: Leading Power Usage as per Load Type

IV. Data Preparation

The data preparation involves transforming the raw data into a format suitable for analysis and modeling. In the context of the Steel Industry Energy Consumption dataset, this preparatory stage is fundamental for ensuring the accuracy, reliability, and effectiveness of subsequent predictive modeling efforts made ahead.

A. Skewness Handling

The initial skewedness for the numerical features in the dataset are first calculated. The data distribution of the Lagging Reactive Power is shown in the figure Fig 8. The log transformation is usually used on the data that is right skewed and since the column is positively skewed as shown in the plot before transformation hence it is right skewed and thus Log Transformation is also used here. Finally the before and after skewness handling transition of the data in the mentioned column is shown side by side with proper comparison in the figure Fig 5.

B. Encoding Categorical feature

Next the categorical columns within the dataset are identified with the help of the R's 'is.character' functionality. Once the categorical columns are identified the categories are encoded with numerical values by using the mutate functionality available in the R package called 'forcats'.

C. Correlation Analysis

Upon checking the correlation matrix represented in the figure Fig 6. The correlation matrix depicts that there isn't any

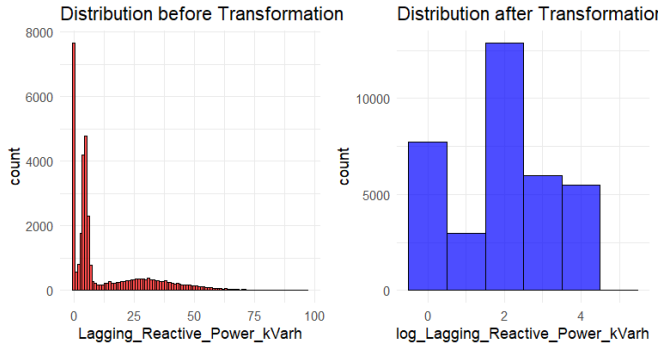

Fig. 5: Log Transformation Before and After Plot

significantly high correlation between the continuous numerical dataset anymore.

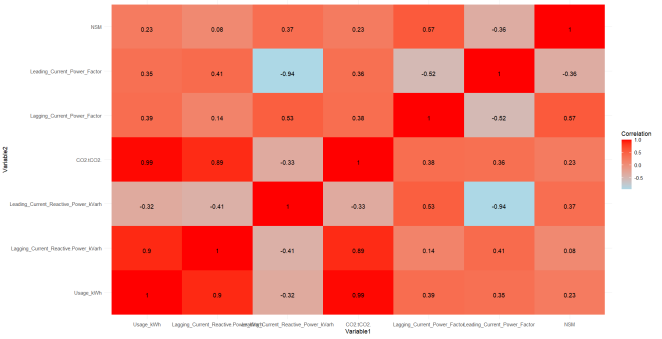

Fig. 6: Correlation Matrix heatmap for Numercial features

D. Principal Component Analysis

The Principal component analysis is done for the whole dataset where the number of components are chosen such that majority or the 95% of the variance within the dataset is captured. Upon reviewing the PCA 'screeplot' the first 7 components seemed to be contributing significantly so the 7 components were chosen. The PCA was calculated by the help of 'prcomp' functionality in R. The PCA screeplot is shown in the figure Fig 7.

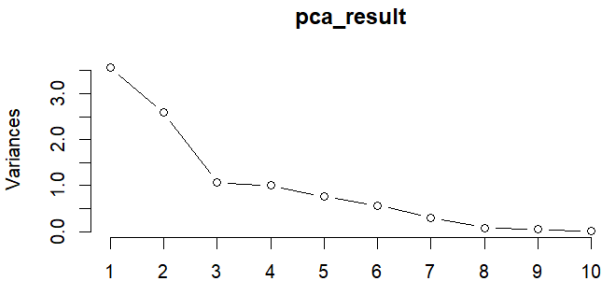

Fig. 7: Scree Plot for Variance explained per PCA

V. Modeling

A. Data Splitting & Training

The Dataset is first split into 2 subsets where all the independednt variables are stored in a variable and the dependent variable is stored in another. The dependent variable 'Load_Type' is contained in a variable train_data whereas rest of the data in the dataset is stored in a subset called test_data. This later helps to divide the steel industry dataset to be finalised that would be used for modeling containing 11 total feature columns and 28032 row observations of data in the training set and 7008 row observations in testing set.

B. Model Selection and Performance

There are two models that are selected here for the purpose of categorizing the Type of Load for the steel industry from the given electrical information. The first model that is being used in Decision Tree. The Ranger implementation of decision trees is an efficient and parallelized version, making it suitable for large datasets to perform classification and regression tasks. The visual representation of a decision tree allows users to easily grasp how input features contribute to predictions. They can capture complex relationships and non-linear patterns, making them suitable for steel industry dataset.
The Second model that is being used is XGBoost, short for eXtreme Gradient Boosting. It is an ensemble learning technique that combines the strengths of multiple weak learners, typically decision trees, to create a robust predictive model. XGBoost often outperforms other machine learning algorithms in terms of predictive accuracy. It is particularly effective in complex tasks and datasets with a large number of features. However for steel industry it can overcome some of the shortcomings of decision tree incase there is any.

VI. Evaluation

The decision tree model over the PCA result dataset gets an accuracy of 89.56% whereas the accruacy for the XGBoost model over the same PCA dataset is just a slightly better and records at 89.86%. Although both the model performs well they perform way better on the normal scaled dataset without PCA where the Decision Tree model's accuracy is measured to be 93.99% along with precision for each class with 0.9878184, 0.8737408 and 0.9008915. The overall F1 score for the model is 92.12%. Finally the XGboost model has an accuracy of 97.78% and precision for all three classes records at 1, 0.9444068, 0.9620187 and the overall F1 score is scored at 96.9% showing barely any class imbalance issues.

VII. Conclusion

In this analysis, the performances of Decision Tree and XGBoost models were compared on a dataset processed with Principal Component Analysis (PCA) and a traditionally scaled dataset. The Decision Tree achieves a commendable accuracy of 93.99%, showcasing balanced precision across different classes. However, XGBoost outperforms with an impressive accuracy of 97.78%, high precision for each class, and an outstanding overall F1 score, suggesting superior generalization capabilities. The choice between the Decision Tree and XGBoost models depends on the specific requirements of the problems. In this case while both models perform reasonably well on the PCA-transformed dataset, the superior performance of XGBoost on the scaled dataset without PCA suggests its effectiveness in capturing intricate patterns and relationships in the original feature space. Additionally, XGBoost's ability to handle class imbalance is a notable advantage.

PART B: Adult Census Dataset

I. Introduction

The Adult Census dataset is a comprehensive collection of demographic and socio-economic attributes, meticulously curated to facilitate machine learning classification projects. Each entry in the dataset represents an individual, providing information on various factors such as age, workclass, education, marital status, occupation, race, and more. The primary goal of this project is to develop a robust machine learning model capable of accurately predicting an individual's income level based on these features. The dataset includes crucial details such as capital gain, capital loss, and hours worked per week, offering insights into financial behaviors and work patterns. Moreover, the native country variable provides a glimpse into the cultural and geographical context of the individuals.

A. Objective

The primary objective of this machine learning classification project is to develop and deploy an accurate predictive model capable of classifying individuals' income levels based on demographic and socio-economic attributes from the Adult Census Dataset. The classification of income is of paramount importance due to its multifaceted significance across various domains. The accurate classification of income levels is crucial for individuals, businesses, and policymakers to make informed economic decisions and provides a foundation for financial planning, budgeting, and investment strategies, contributing to overall economic stability. The insights derived from this classification endeavor to inform decisions that contribute to economic stability, societal equity, and the well-being of individuals and communities.

II. Related Work

The foundational resource for this research stems from the "UCI Machine Learning Repository" [Dua and Graff, 2019], which hosts the Adult dataset, forming a cornerstone for numerous studies in machine learning and data analysis. Building on this, "The Elements of Statistical Learning" by Hastie, Tibshirani, and Friedman [2009] provides a comprehensive textbook covering statistical learning principles, offering a theoretical framework applicable to machine learning techniques deployed on census datasets. In the realm of predictive modeling, Johnson et al.'s study on a pediatric emergency department [2018] provides insights into methodologies that may find application in the analysis of census data. Addressing income prediction, Bhavsar and Shah's work [2019] delves into machine learning and deep learning approaches, offering valuable perspectives on predicting income-related variables from census data.

age	workclass	fnlwgt	...	native-country	income
39	Private	77516	...	United-States	<=50K
50	Self-emp-not-inc	83311	...	United-States	<=50K
38	Private	215646	...	United-States	<=50K
53	Private	234721	...	United-States	<=50K
28	Private	338409	...	Cuba	<=50K

TABLE II: Adult Census Income Dataset - Sample

III. Data Mining Methodology & EDA

The Knowledge Discovery within the datasets are done with the help of vast number of packages available in R programming language. The dataset is fetched from UCI machine learning dataset repository online.

A. Dataset Summary

The Adult Census Income Dataset, sourced from the census bureau database and donated by Ronny Kohavi and Barry Becker, encompasses 48,842 instances with a combination of continuous and discrete features. The prediction task involves determining whether an individual earns over 50K a year. Notably, the dataset includes demographic factors such as age, education, occupation, and native country. The dataset is shown in the table II. The dataset when viewed and investigated, it was found that some of the observations for certain columns such as 'native country', 'occupation' and 'workclass' have missing values that are replaced with the symbol '?'. So with the help of R's basic dataframe manipulation techniques the question marks are replaced with 'NA' that certainly denotes missing values. Upon checking the dataframe's missing values now it shows that there are 1836 missing values for the column workclass, 1843 missing values for the column 'occupation' and 583 missing values for 'native country'.

B. Missing Data

Once the miissing values are identified they need to be handled. The number of missing vallues are identified above and now to handle them the native country column is picked first. Since this column doesn't provide too much of a crucial information and has very negligible number of null values when compared to the dataset size, the null values are instantly dropped.
Next the columns 'occupation' and 'workclass' are handled by replacing the null values with the mode values in the columns.

C. Categorical Features Exploration

Next as per the dataset description from the source and also checking the dataset when viewed, it is found that the dataset contains quite a few categorical columns. The categorical columns name were "workclass", "education", "marital_status", "occupation", "relationship", "race", "sex", "native_country" and "income". All the categorical columns are encoded with the help of Label Encoding feature available in R where the encoding is done to all the unique elements in the columns. There are total 9 categorical columns within the dataset.

D. Visual Exploration

Within the visual exploration at first the countplots for Age and income are visualised. The plots are shown in the figure Fig 8 and 9. Upon checking the income countplot it is evident that

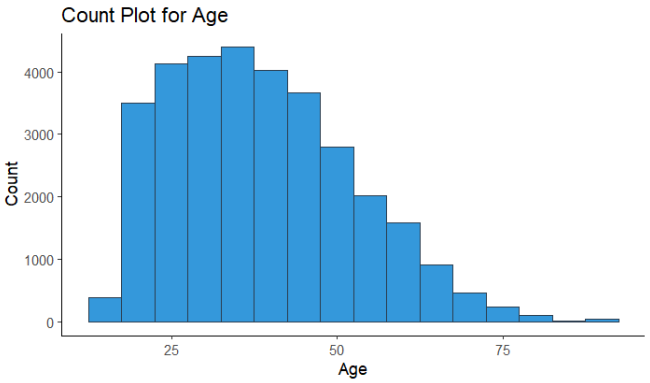

Fig. 8: Data distribution of Age Feature

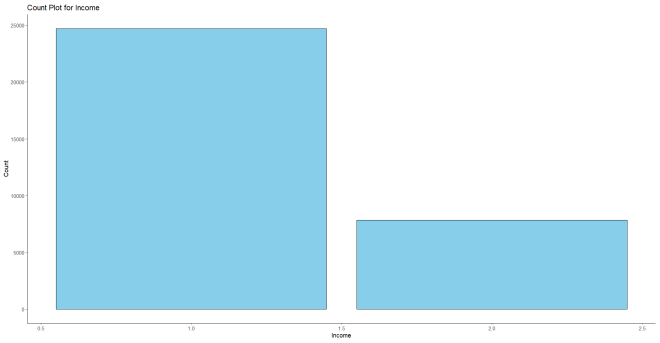

Fig. 9: Countplot for Target(Income) Variable

the data has quite a bit of data imbalance issue. The classes are divided into a 65 and 35 percentage division. The age however has a good spread and a bit of right skewed. The most observations are recorded between 25 to 50.

E. Correlation Matrix

Within the EDA the correlation matrix provides valuable insights into the relationships between different variables in the dataset. Positive correlations close to 1 indicate a strong linear relationship between certain pairs of variables. However upon checking the heatmap there were some columns that had barely any correlation with the target column. That is determined by checking the values closer to 0 signifying weak or no linear correlation. The overall correaltion matrix is shown in the figure Fig 10. Next there are some of the visualizations in between certain features and the target variable. The first plot represents the division of the incomes based on the categories in occupation types. The plot is shown in figure Fig 11. It can be concluded that mostly within the occupations, the no. of people with income

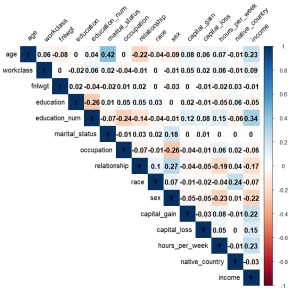

Fig. 10: Correlation Matrix Heatmap

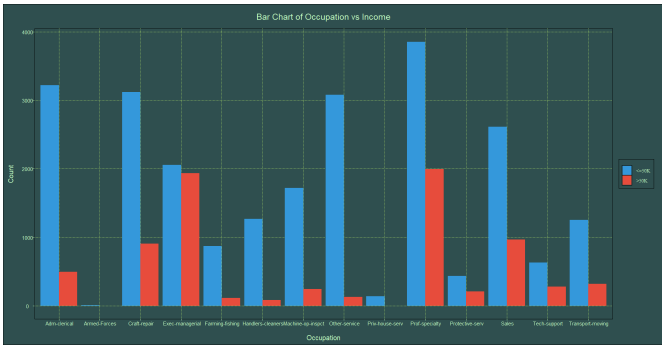

Fig. 11: Barplot for Occupation as per Income

less than 50 k is 3 times more than the no.of people with more than 50k earning.

The next bar plot explores the Workclass vs the Income. The income distribution as per the categories in the feature workclass are shown in the figure Fig 12. It is quite evident that self employed people have a larger bracket of people with an earning greater than 50k.

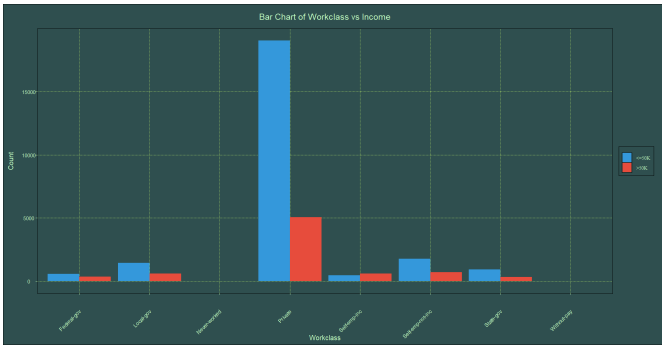

Fig. 12: Barplot for Workclass as per Income

Moving forward the next plot has the income portrayal according to the Marital status of the observed individuals. The figure Fig 13 shows the spread. The plot shows that the married people are more in the 50K plus bracket, whereas people who never married largely lie in the group of people earning less than 50k. Next the boxplots for Age and Hours per Week are plotted to check for outliers present within the dataset. The figures Fig 14

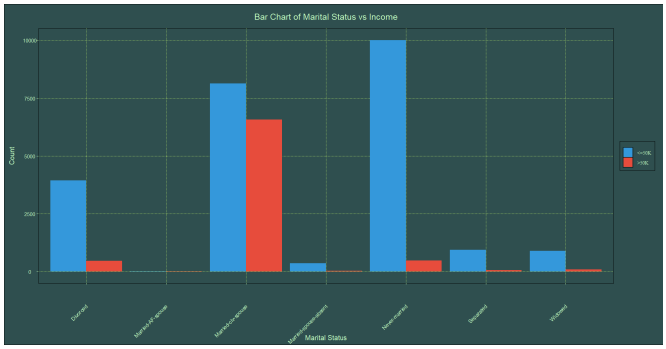

Fig. 13: Barplot for Marital Status as per Income

and Fig 15 show the boxplots. It can be seen that the features exhibit quite a bit of outliers.

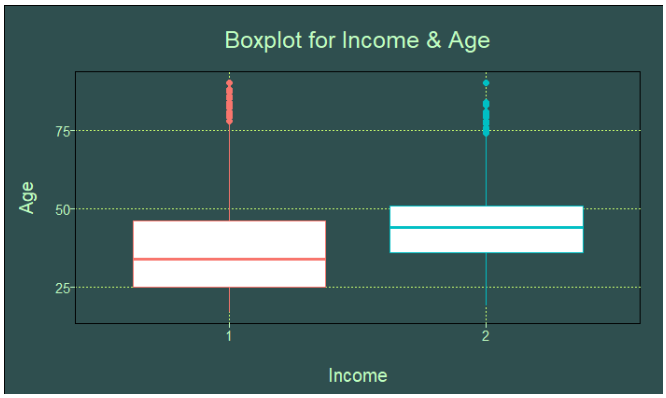

Fig. 14: Boxplot for Age

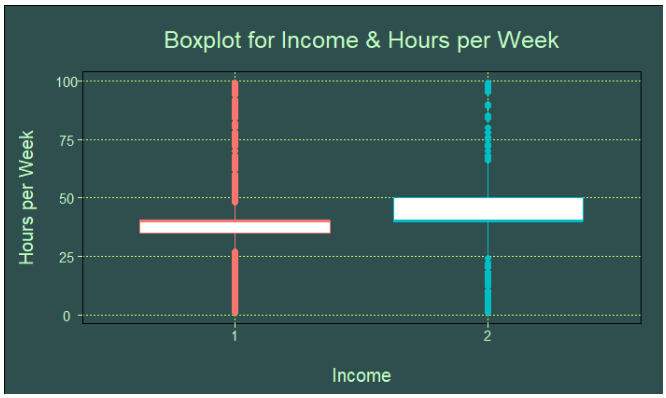

Fig. 15: Boxplot for Hours Per Week

IV. Data Preparation

The data preparation involves transforming the raw data into a format suitable for analysis and modeling. In the context of the adult census dataset, this preparatory stage is fundamental for ensuring the accuracy, reliability, and effectiveness of subsequent predictive modeling efforts made ahead.

A. Outlier Handling

The Boxplots shows quite a fair share of outliers within the plots. Firstly the outliers are identified by the help of z-score method. The threshold is set as 3. The outliers are then handled one by one for each and every numeric continuous coolumns. The identified observations that contains the outliers are then removed from the dataset.

B. Scaling Dataset

The outlier handled data is then scaled and normalized using the min max normalization method. Upon encoding the dataset it is also found that the target variable has 3 level or categories and are encoded as 1, 2 and 3. However some models prefer the encoding to be starting from 0. So the values for the income column are decreased by 1 unit.

V. Modeling

A. Data Splitting & Training

The Dataset is first split into 2 subsets where all the indepenednt variables are stored in a variable and the dependent variable is stored in another. The dependent variable 'income' is contained in a variable train_data whereas rest of the data in the dataset is stored in a subset called test_data. This later helps to divide the adult dataset to be finalised that would be used for modeling containing 14 total feature columns and 30131 total observations of data, out of which 21092 observations are in the training set and 9039 observations are in the testing set.

B. Model Selection and Performance

There are two models that are selected here for the purpose of categorizing the income for the adult census data from the given adult information. The first model that is being used in Logistic Regression a statistical method widely used for binary classification problems, where the target variable is categorical with two classes. The logistic function, also known as the sigmoid function, is integral to this algorithm. The adult dataset also has the target column in a binary format so logistic regression suits very well for the purpose of classification in this scenario.

The Second model that is being used is Random Forest, which is an ensemble learning method that leverages the power of multiple decision trees to improve predictive accuracy and control overfitting. It is a versatile algorithm capable of handling both classification and regression tasks. By combining the predictions of several trees, Random Forest offers robust performance and resilience to noisy or complex datasets.

VI. Evaluation

The Logistic Regression model achieved an accuracy of 82.71%, indicating its ability to correctly classify instances into the target classes. Precision, which measures the accuracy of positive predictions, is 69.77%, indicating that when the model predicts an individual earns more than 50K, it is correct 69.77% of the time. Recall, representing the ability to correctly identify positive instances, stands at 39.98%. The F1 Score, a balanced metric of precision and recall, is 0.5083. The model demonstrates a higher accuracy for Class 0 (income ¡= 50K) with an accuracy of 95.01%, but a lower accuracy for Class 1 (income ¿ 50K) at 39.98%

VII. Conclusion

The Random Forest model emerges as the top performer, offering the highest overall accuracy and a balanced performance for both classes. Logistic Regression, while demonstrating reasonable accuracy, lags behind in precision and recall compared to Random Forest. Random Forest stands out as a robust choice for this classification task, providing high accuracy and balanced performance. Fine-tuning hyperparameters and exploring feature engineering may further enhance model performance.

PART C: Bank Marketing Dataset

I. Introduction

The dataset under consideration pertains to the direct marketing campaigns conducted by a Portuguese banking institution. These campaigns, which primarily involved phone calls, were executed to promote a term deposit subscription. The dataset encompasses various features related to the clients, such as age, job, marital status, education, and financial information like balance. The objective is to predict whether a client will subscribe to the term deposit ('yes') or not ('no'). The classification task holds significance for optimizing marketing strategies and resource allocation.

A. Objective

The primary objective of this project is to build predictive models that can accurately forecast whether a client will subscribe to the term deposit. By leveraging statistical analysis, exploratory data analysis (EDA), the project aims to vevelop machine learning models to predict the likelihood of a client subscribing to the term deposit based on various client attributes and interaction history. Alongside it also compares the performance of those models to identify the most effective approach for subscriptions prediction.

The project's outcome is expected to provide valuable insights for the bank, enabling them to tailor their marketing efforts more effectively and improve the success rate of term deposit subscriptions.

II. Related Work

The foundational resource for our exploration into the Bank Marketing dataset is the work by Moro, Cortez, and Rita [2014], titled "A Data-Driven Approach to Predict the Success of Bank Telemarketing." This seminal paper introduced the dataset, providing key insights into its features and the predictive modeling task. Extending our focus to the application of machine learning algorithms, Ahmed et al. [2018] conducted a comprehensive comparative analysis titled "Comparative Analysis of Machine Learning Algorithms for Bank Telemarketing." This study evaluates various algorithms, including K-Nearest Neighbors (KNN), shedding light on their performance in the context of bank telemarketing.

III. Data Mining Methodology & EDA

The Knowledge Discovery within the datasets are done with the help of vast number of packages available in R programming language. The dataset is fetched from UCI machine learning dataset repository online.

A. Dataset Summary

The dataset comprises information on 45,211 client interactions, each described by 17 variables. Key client attributes include age, job type, marital status, education, default status, balance, housing loan status, and personal loan status. Additionally, details about the mode of contact, the day of the week, and the month of the interaction are available. The duration of the call, the number of contacts performed during this campaign (campaign), and historical information such as the number of days since the client was last contacted (pdays) are also part of the dataset. The features have some certain trends some of which are like that the clients' ages range from 18 to 95, with a mean age of approximately 41, Job types vary across categories such as blue-collar and many other identifiable trends are present in the dataset.

B. Correlation Analysis

To understand the relationships between numerical variables, a correlation matrix was computed. The matrix helps identify potential multicollinearity and offers insights into variables that might influence the target variable. The figure Fig 16 depicts the correlation matrix.

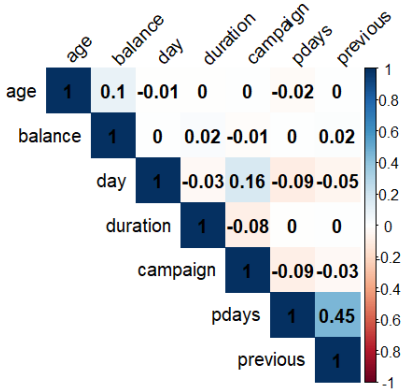

Fig. 16: Correlation Matrix in Bank Dataset

The dataset when viewed and investigated, it was found that some of the observations for certain columns such as 'native country', 'occupation' and 'workclass' have missing values that are replaced with the symbol '?'. So with the help of R's basic dataframe manipulation techniques the question marks are replaced with 'NA' that certainly denotes missing values. Upon checking the dataframe's missing values now it shows that there are 1836 missing values for the column workclass, 1843 missing values for the column 'occupation' and 583 missing values for 'native country'.

C. Missing Data

The dataset has been thoroughly examined for missing values, and the results indicate that there are no null values present in any of the columns. The absence of missing values simplifies the data preprocessing phase, eliminating the need for imputation or deletion of records with missing information.

D. Categorical Features Exploration:

The dataset contains several categorical columns that provide insights into the demographic and contact information of the clients. The categorical columns include: Job, Marital status, Education level, Default status, Housing loan status, Personal loan status, Contact communication type, Month of the last contact, Outcome of the previous marketing campaign, Subscription to a term deposit (target variable)

E. Visual Exploration:

Within the visual exploration at first to better understand the distribution of numeric features, histograms were created for each numeric column. These visualizations provide a clear overview of the data distribution, allowing for the identification of patterns and potential outliers. The histograms showcase the frequency distributions in the figures Fig 17,18 and 19.

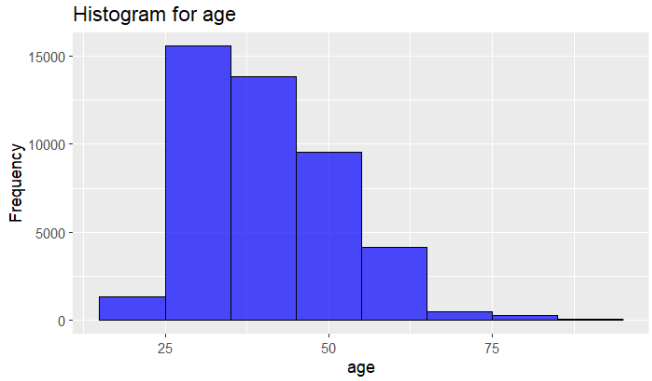

Fig. 17: Data Distribution for Age

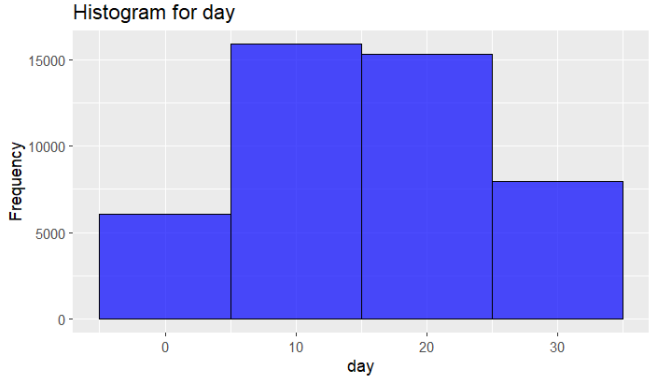

Fig. 18: Data Distribution for Day

Upon checking the plots age is mostly distributed from 25 to 60 however the total data has wider age range. The duration

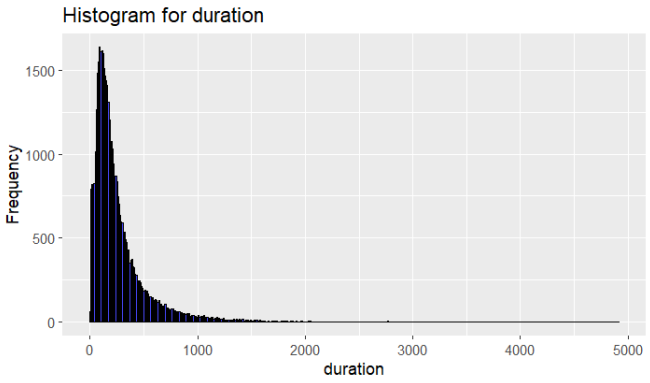

Fig. 19: Data Distribution for Duration

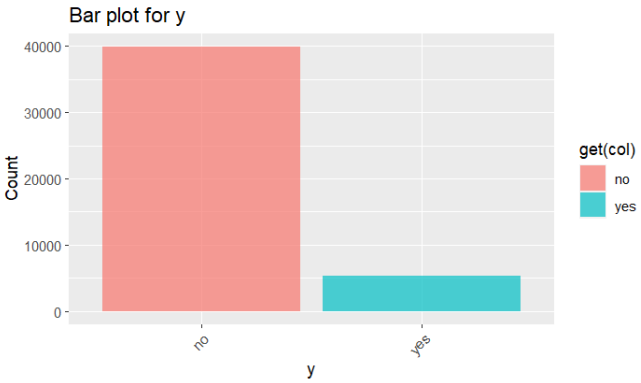

Fig. 22: Target Variable Countplot

distribution is right skewed data. The categorical columns however are checked for their categories by the help of bar charts. The bar charts are represented in the figures Fig 20, 21 and 22 . Boxplots serve as powerful tools in visually summarizing

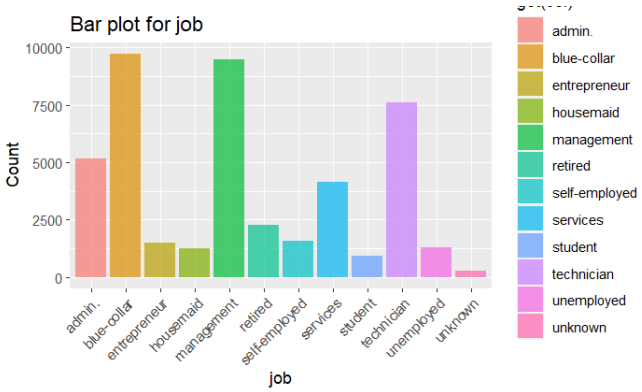

Fig. 20: Bar Plot for Job Types

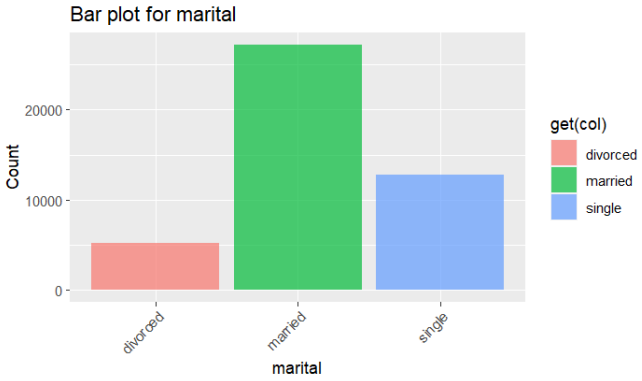

Fig. 21: Bar Plot for Marital State

the distribution of numerical variables in the bank dataset. The 'balance' variable boxplot illustrates a right-skewed distribution, suggesting that a majority of clients have lower balances, with a few having significantly higher balances. Outliers are present on the upper end, indicating a small proportion of clients with exceptionally high balances. The boxplot for the 'duration' variable, representing the duration of the last contact, shows a positively skewed distribution.

IV. Data Preparation

The data preparation involves transforming the raw data into a format suitable for analysis and modeling. In the context of the bank marketing dataset, this preparatory stage is fundamental for ensuring the accuracy and reliability. Most of the preparation is shown in the EDA part however the outliers are handled at this point.

Outlier Handling: The Boxplots shows quite a fair share of outliers within the plots. Firstly the outliers are identified by the help of checking the upper and lower bound after calculating the inter quartile range. The data present within the upper and lower bound of the IQR calculated is retained and the rest is discarded as they are supposedly the outliers. The Boxplots after calculating the outliers are plotted in the figures Fig through Fig

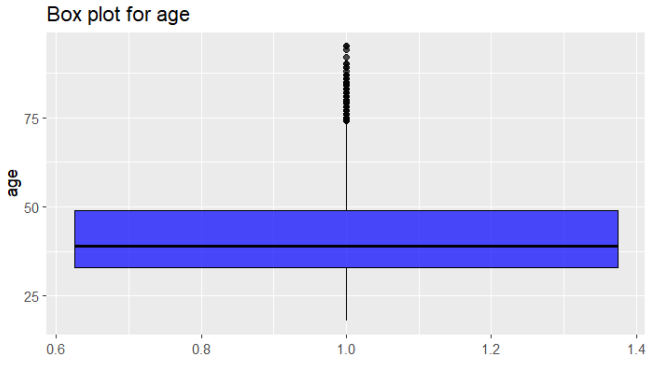

Fig. 23: Box Plot for Age

V. Modeling

Data Splitting & Training:
The dataset is split in the same way it has been done for the other two datasets above in a 80 to 20 percent format.

Model Selection and Performance:
There are two models that are selected here for the purpose

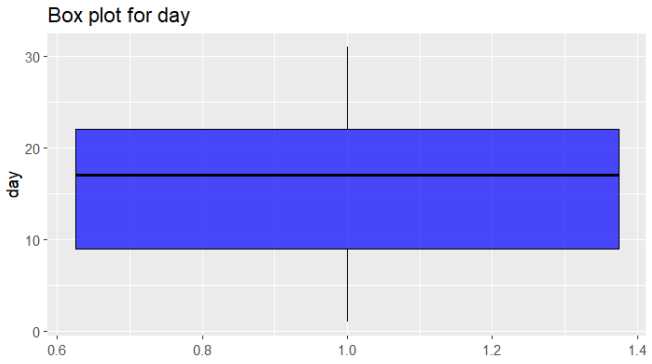

Fig. 24: Box plot for Day

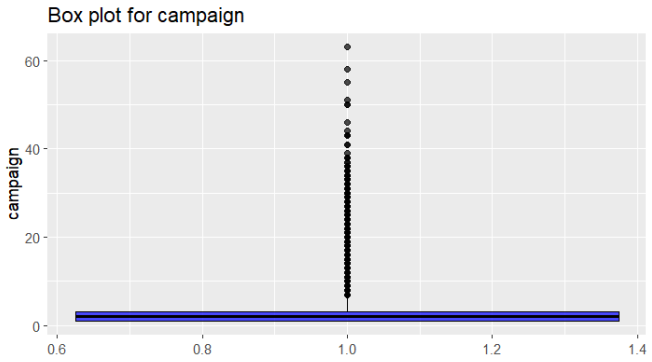

Fig. 25: Box Plot for Campaign

of categorizing the data deciding wheather or not the client subscribed to a term deposit. The first model that is being used is K-Nearest Neighbors (KNN), which is a non-parametric and instance-based machine learning algorithm used for classification tasks. In the context of the bank dataset, KNN aims to classify clients based on their similarity to others in the dataset. The algorithm assigns a new data point to the majority class among its k-nearest neighbors.

The other model that is being used here is XGBoost, short for Extreme Gradient Boosting, is an ensemble learning algorithm that combines the predictions of multiple weak learners (decision trees) to create a robust model. XGBoost's strength lies in its ability to handle complex relationships within data and mitigate overfitting.

VI. Evaluation

The K-Nearest Neighbors (KNN) model achieved an accuracy of 90.50%, with a precision of 46.51%, recall of 23.60%, and an F1 score of 0.3131. Despite high accuracy, the model suffered from class imbalance, leading to significantly lower recall and F1 scores, indicating a potential to miss actual positive instances. Further optimization, including tuning 'k' and exploring feature engineering, may enhance performance.

In comparison, the XGBoost model showed an accuracy of 91.69%, precision of 62.19%, recall of 41.12%, and an F1 score of 0.4951. While facing challenges in identifying positive instances, XGBoost demonstrated improvements over KNN, with higher precision and recall, suggesting better performance in detecting actual positive instances.

VII. Conclusion

The analysis revealed nuanced insights into the strengths and weaknesses of each model. KNN, characterized by its simplicity, demonstrated a reasonable accuracy of 90.50%, yet its precision and recall metrics indicated limitations, particularly in dealing with imbalanced datasets. On the other hand, XGBoost, an ensemble learning approach, outperformed KNN on all fronts. With an accuracy of 91.69%, significantly improved precision (62.19%), and enhanced recall (41.12%), XGBoost exhibited a more robust capability to identify positive instances while maintaining a low rate of false positives. The comparative analysis underscores the superiority of XGBoost for the given task, providing a concrete recommendation for its adoption in predicting term deposit subscriptions. The research not only contributes to the specific domain of banking but also underscores the broader significance of selecting appropriate machine learning algorithms tailored to the characteristics of the dataset at hand. The results encourage further exploration, including hyperparameter tuning and addressing class imbalance, to unlock the full potential of predictive modeling in banking and related domains.

Acknowledgments

The completion of this project has been made possible through the support from Professor Musfira Jilani, for her guidance, advice and critical feedback.

References

[1] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[2] L. Breiman, "Random Forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.
[3] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.
[4] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine learning, vol. 63, no. 1, pp. 3-42, 2006.
[5] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," Springer, 2009.
[6] D. R. Cox, "The regression analysis of binary sequences," Journal of the Royal Statistical Society: Series B (Methodological), vol. 20, no. 2, pp. 215-242, 1958.
[7] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.
[8] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," The American Statistician, vol. 46, no. 3, pp. 175-185, 1992.
[9] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.
[10] Fan, C., Xiao, F., & Wang, S. "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques" , 2014
[11] Auguie, B. "Gridextra": "Miscellaneous functions for "grid graphics". R Package Version, 2(1), 1–9, 2017
[12] Arnold, J. B. "Ggthemes": Extra themes, scales and geoms for "ggplot2". R Package Version, 3(0), 1–284, 2017