National College of Ireland

Project Submission Sheet

Student Name: Zuu Zuu Kyaw Shwe

Student ID: 24106585

Programme: MSCDAD_C **Year:** 2024

Module: Data Mining & Machine Learning

Lecturer: SHERESH ZAHOOR

Submission Due Date: 15th Dec 2024

Project Title: Continuous Assessment

Word Count: 5722

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Zuu

Date: 15th Dec 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Data Mining & Machine Learning

Project Final Report

Zuu Zuu Kyaw Shwe
School of Computing (MSCDAD-C)
National College of Ireland
Dublin, Ireland
x24106585@student.ncirl.ie

***Abstract*—This project incorporates machine learning in three different aspects which are airline passenger satisfaction, energy appliance usage, and Twitter sentiment analysis. The study aimed to understand the data by preprocessing, feature engineering, and evaluating models using the CRISP-DM approach, hoping to bring forth some useful optimizations. In the analysis of satisfaction among passengers in the airline industry, delays and service quality emerged as the main factors with XGBClassifier logging in a good amount of accuracy. An analysis of energy consumption pinpointed temperature and time as key indicators whereas sentiment classification on Twitter indicated that text preprocessing is an essential step in the modeling process while ensemble models did much better than simpler ones. The results emphasize the ubiquity of machine learning in building models directed at solving different problems based on the defined equilibrium between the complexity and interpretability of the model. In subsequent works, it will be possible to use deep learning and integration of external data sources to improve the quality of forecasts and their application.**

***Index Terms*—Machine Learning, Airline Passenger Satisfaction, Energy Consumption Prediction, Sentiment Analysis, Ensemble Models, CRISP-DM, Preprocessing, Feature Engineering**

I. INTRODUCTION

The crucial role of data analysis in solving different problems arising from improving customer experience to enhancing energy conservation has accelerated over time. The emergence of huge volumes of data and the improvement of machine learning methods have made the application of systematic approaches to solving such tasks quite popular. This project involves three separate but interlinked investigations: airline passenger satisfaction, energy appliance usage, and Twitter sentiment classification, which are all examples of solutions to data mining problems.

The first study examines airline passenger satisfaction with the aim to identify the factors that impact the customer experience and evaluate the predictive models that have been developed for satisfaction levels. It is essential that these elements are recognized so that better service quality is rendered and customers remain loyal to companies in the aviation sector. The second component explores energy consumption by appliances where the aim is to give an account of energy usage and the factors that drive that use. The outcomes of this research assist energy conservation measures and the management of resources sustainably. The Twitter sentiment analysis finally focuses on the task of classifying Twitter sentiment where there is a great prevalence of non-standardized language, informal communication, and slang words or phrases, and multiple sentiment expressions when used on Twitter.

Extracting knowledge from complex datasets and resolving domain-oriented issues is within the scope of this project, which seeks in its totality to demonstrate machine learning's utility in real-life applications. This work spans the rigorous data preparation stage through to model development and evaluation in all studies, CRISP-DM's framework for a systematic approach. The research problems are posed in determining the major contributors in predicting the outcome, measuring the efficacy of some new forecasting machine learning models, and improving the methods in terms of understanding and their application to other cases.

The remainder of this document is organized as follows: the Data Mining Methodologies section describes the detailed steps followed in every single analysis and also presents the analysis of model performance and the resulting effects. The Related Work section provides an overview of where the project rests with respect to already existing studies. Finally, in the final chapter, outline and summarize the main goals of the study and suggest future directions for the growth and continuation of the research that has been initiated.

II. RELATED WORK

A. Airline Passenger Satisfaction Analysis

Airline companies are increasingly being analyzed for satisfaction through machine learning and its predictive power on key determinants. This section reviews and critiques relevant studies, linking them to this analysis.

Hibović et al. employed KNN, Naïve Bayes, and Random Forest classifiers for classifying passengers using the dataset of passenger satisfaction and stated Naïve Bayes as the best classifier, where Boruta algorithm was used to select variables and missing values were filled. Nevertheless, their work has not attempted to investigate the interdependency between features like Arrival and Departure Delays [1]. On the other hand, in our work, we used ensemble models – XGBoost and CatBoost – which yielded 95.9% accuracy and 99.42% AUC score, with Entertainment and Seat Comfort being the most important aspects of satisfaction. Adding Boruta based variable selection might make the performance of our model even better.

Shu analyzed the impact of flight delays and cancellations by using machine learning in a form of algorithms, arguing their importance from the perspective of satisfaction. The aforementioned features' significance in their study such as Departure Delay and Arrival Delay are in align with our studies that say operational delays greatly undermine satisfaction. However, their research did not combine these operational factors with a comprehensive model of satisfaction prediction [2].

Tan used Bidirectional LSTM models to classify the sentiments on the scribbled reviews of airline passengers and attained accuracy levels of 91.27%, reasoning out some of the negative comments to be poor customer service and late flights. Although the importance of unstructured data in satisfaction research is highlighted in their work, combining these insights with structured data—like evaluations from Inflight Entertainment—could improve a more thorough picture of passenger contentment [3].

Zhao investigated the use of fuzzy comprehensive evaluation methods when evaluating customer satisfaction in terms of service quality, punctuality or comfort. This procedure deals with qualitative aspects in a more organized manner which also assists in tackling the problem of assessing various aspects of satisfaction [4]. Although our focus in this paper was on the application of various machine learning techniques for prediction, fuzzy evaluation methods such as the one proposed by Zhao might aid in the quantitative models by adding some qualitative input variables, leading to enhanced understanding of passenger satisfaction. Such an integration could also make results more comprehensible as well as improve the overall analysis.

B. Energy Appliances Analysis

The patterns of use of energy and the economy of energy also make the forecasting of energy demand an important task. A number of studies have looked into the prospects of employing machine learning methods that have been complemented with feature selection for successful forecasting.

It was noted by El-Gohary and Amasyali that Random Forest and Gradient Boosting represents effective means of dealing with non-linearities and feature importance ranking and estimation tasks. Further, they pointed out important deficiencies, such as the lack of proper investigations into long-term and residential energy forecasting and modeling that includes the behavior of building occupants [5].This is in line with our research as well, since we employ data-driven regression models to estimate the consumption of energy by appliances, however future work may be enhanced by taking into account changes in occupant behavior and long term trends of consumption patterns in the forecast.

Ahmad and Chen looked at applications of machine learning in energy systems and outlined the significance of supervised models such as support vector machines and neural networks. It was established that these models are applicable in real-time plain, for instance, in load demand and renewable energy forecasting. Such results also emphasize the flexibility of machine learning methods in energy prediction that correlate well with our interest in using regression techniques for appliance energy forecasting [6].

Liu et al. adopted Random Forest (RF) and Gradient Boosting Regression (GBR) as their machine learning models to forecast building energy use while emphasizing feature selection to avoid high dimensionality and preserve accuracy. Their use of RF and GBR is also in agreement with our use of ensemble methods that aim to improve the prediction regardless of the existing non-linear relationships. The methods proposed by them for feature engineering and features ranking make it possible to appreciate the very essence of the application of these techniques in energy forecasting and such principles are also part of our efforts in regression modeling of appliance energy prediction. In the same way, both studies reveal the robustness and usefulness of ensemble models for analyzing complex energy consumption data in relation to the forecasting target variables [7].

According to Mohapatra et al. energy consumption forecasting using deep learning methods consists of Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). They showed that GRU appears to be more beneficial than LSTM as regards to lowering model loss and improving accuracy which makes it more suitable for time-series data [8]. This however corresponds to the orientation of our research which also dealt with regression models but due to the complexity of the task settled for a more basic and interpretable approach. In our case, the future work can integrate GRU or LSTM models into our model which would allow for better temporal features in energy consumption to be utilized, thus extending our ensemble regression methods.

C. Twitter Sentiment Analysis

We have examined the harsh reality which is the availability of emotionally biased opinions in the form of text due to sentiment analysis on social media platforms such as Twitter. A number of conventional approaches to automated sentiment extraction have originally emerged and have been evolving throughout the years.

Zhao et al. performed a comprehensive assessment of the effects of various preprocessing methods, such as acronyms expansion, negation and stopword removal, on sentiment analysis of Twitter data. Their findings reveal that practices like acronym expansion as well as negation replacement not only enhance model performance, but also improve the feature representation while lessening the amount of noise. Likewise, our study utilized efficient preprocessing techniques such as stop words and lemmatization, to enhance the quality of the data. While Zhao et al. dealt mainly with the effects of preprocessing, this work builds on this by using more complex feature extraction processes such as TF-IDF for increased classification precision [9].

Sarlan et al. explored lexicon-based two sentiment analysis techniques to deal with emoticons, slang, and abbreviations language in Twitter while highlighting Twitter language as an important challenge for sentiment analysis. However, they noted that the focus on such lexicons only partially resolved and provided insufficient means to cope with the ever-changing forms of social media language [10]. In contrast, our study employs TF-IDF vectorization, which models the weight and relevance of words in context, overcoming the rigidity of static lexicons and enhancing sentiment prediction accuracy.

Khurana et al. compared issues with Naïve Bayes, SVM and Random Forest for sentiment classification, namely that Naïve Bayes was most effective on simpler datasets while SVM showed most promise for tackling complex high dimensional data. Random forests, however, excelled in their resilience to various architectures. Their focus on reducing weaknesses of the model coincides with our application of the XGBoost and Extra Trees ensemble methods to improve prediction and control complexity at the same time [11].

Mandloi and Patel researched how the Naïve Bayes, SVM, and Maximum Entropy models work for sentiment classification tasks and in this case they worked with Twitter data. Their work also offered a careful consideration of the accuracy versus time complexity tradeoffs inventing Naïve Bayes as the least complex algorithm, but claiming that Maximum Entropy yielded better results in more complex settings. In their work however they found that SVM was not as precise but was able to work more reliably in high dimensional space [12].The study fits with our analysis as it also points to the limitations of natural models such as the Naïve Bayes when analyzing finer datasets thus justifying our decision to employ ensemble methods so as to mitigate this to some extent and increase the predictive capabilities.

III. Data Mining Methodologies

A. Airline Passenger Satisfaction Analysis

1) Business Understanding: The main objective was to assess the critical factors that influence passengers and evaluate the performance of machine learning models in making predictions. Such factors are crucial for enhancing airline services and retaining customers. The target variable, Satisfaction, measured the degree of satisfaction of passengers, classifying them as either Satisfied for those responding positively or Neutral/Dissatisfied for the rest.

2) Data Understanding: The dataset contained 24 columns and 129,880 rows, consisting of categorical and numerical data. Important variables included demographic data (e.g., Gender, Age), travel characteristics (e.g., Class, Flight Distance), and service quality ratings (e.g., Entertainment, Seat Comfort). Exploratory Data Analysis (EDA) was then performed to examine the data to find patterns and relationships within the variables:

***Target Variable Distribution*:**

Satisfaction Balance: The dependent variable, Satisfaction, was relatively balanced, though there was a small skew towards satisfied passengers (approximately a 9.4% difference).

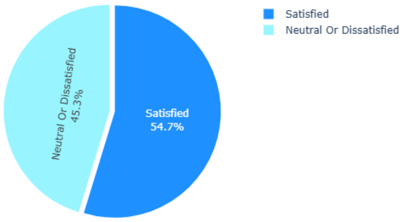


Fig. 1: Target Variable Distribution

***Identification of Outliers*:**

Outliers Analysis: There were outliers in the key columns of Flight_Distance, Departure_Delay_Minutes, and Arrival_Delay_Minutes and they were maintained as they do not accurately reflect the data if they are removed or filled in.

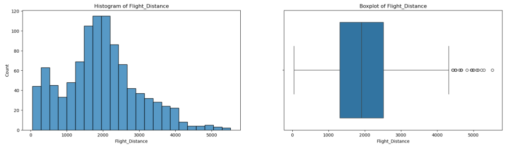


Fig. 2: Histogram & Boxplot of Flight Distance

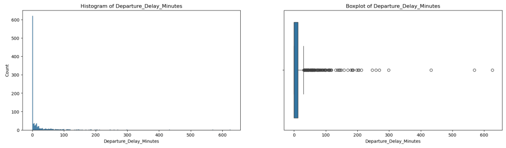


Fig. 3: Histogram & Boxplot of Departure Delay

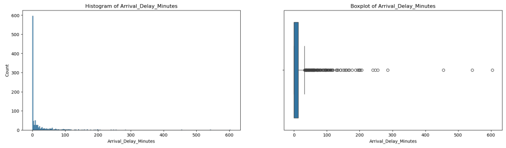


Fig. 4: Histogram & Boxplot of Arrival Delay

***Demographic and Behavioral Insights*:**

Gender: Gender did not significantly influence satisfaction. The number of men and women in this sample was almost the same.

Customer Type: The airline had a significant number of loyal customers, many of whom are satisfied; however, a notable proportion of these loyal customers were neutral or dissatisfied.

***Feature Relationships*:**

Delay Relationship: Departure delays strongly impacted arrival delays, confirming their interdependence. Passenger dissatisfaction increased with longer delays, highlighting the importance of minimizing departure delays to maintain customer satisfaction.

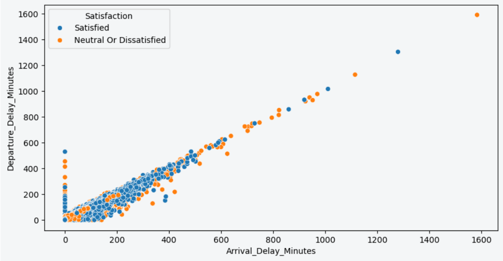


Fig. 5: Correlation Between Arrival and Departure Delay

***Service-Related Insights*:**

Highly Rated Features: Entertainment, Online Booking Service, and Onboard Service were consistently rated higher by satisfied passengers.

Less Impactful Features: Gate Location and Check-in Service appeared to have limited influence on customer satisfaction based on the analysis.

3) Data Preparation: Comprehensive preprocessing steps were performed to ensure the dataset's suitability for machine learning tasks:

Data Cleaning: The ID column was removed because it was not pertinent to the analysis. Renamed columns for clarity and consistency. No duplicates were found in the dataset. Lastly, standardized categorical values to resolve capitalization inconsistencies.

Handling Missing Values: There were 393 missing values in Arrival_Delay_Minutes column and it was imputed using the median to maintain data integrity.

Feature Transformation: To adjust non-normal distributions, a log of variables (Flight_Distance, Departure_Delay_Minutes, Arrival_Delay_Minutes) was taken. Multi-class categorical (Class) and Binary variables such as gender and customer type were transformed using one hot and label encoding, respectively.

Feature Scaling: StandardScaler was used to normalize numerical features.

Feature Selection: The models for feature importance included a Random forest model with variables Entertainment and Seat_Comfort being significant while Class_Eco Plus was omitted as it was not significant.

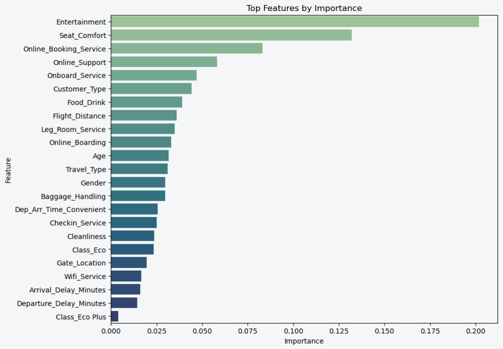


Fig. 6: Top Features by Importance

Correlation analysis revealed redundancy between Departure_Delay_Minutes and Arrival_Delay_Minutes. The former was dropped to reduce dimensionality.

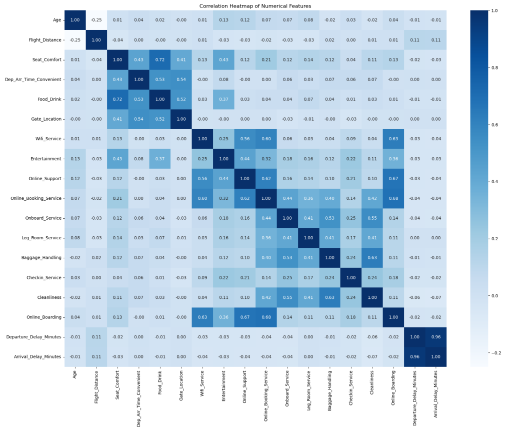
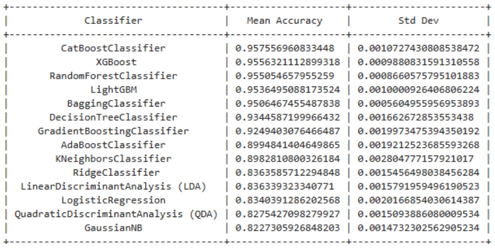


Fig. 7: Correlation Heatmap of Numerical Features

4) Modeling & Evaluation: The modeling and evaluation procedure involved two phases; the first one involved benchmarking through cross-validation and the second one involved a comprehensive design of selected models. In the exploratory phase, CV was implemented over a wide range of common algorithms because we were not sure which algorithm to choose and compare when training a machine learning model.

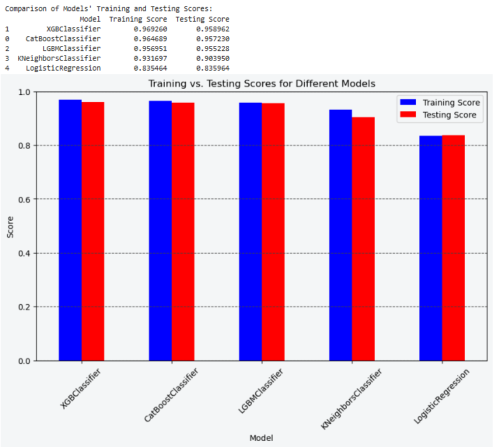


```
+----------------------------------+------------------+----------------------+
|            Classifier            |  Mean Accuracy   |       Std Dev        |
+----------------------------------+------------------+----------------------+
|        CatBoostClassifier        | 0.957556960833448 | 0.00107274308085384 72 |
|             XGBoost              | 0.9556321112899318 | 0.0009880831591310558 |
|      RandomForestClassifier      | 0.955054657955259 | 0.0008660575795101883 |
|             LightGBM             | 0.9536495088173524 | 0.0010000926406806224 |
|        BaggingClassifier         | 0.9506467455487838 | 0.0005604955956953893 |
|      DecisionTreeClassifier      | 0.9344587199966432 | 0.001662672853553438 |
|    GradientBoostingClassifier    | 0.9249403076466487 | 0.0019973475394350192 |
|        AdaBoostClassifier        | 0.8994841404649865 | 0.0019212523685593268 |
|       KNeighborsClassifier       | 0.8982810800326184 | 0.002804777157921017 |
|         RidgeClassifier          | 0.8363585712294848 | 0.0015456498038456284 |
|  LinearDiscriminantAnalysis (LDA) | 0.836339323340771 | 0.0015791959496190523 |
|        LogisticRegression        | 0.8340391286202568 | 0.0020166854030614387 |
| QuadraticDiscriminantAnalysis (QDA) | 0.8275427098279927 | 0.0015093886080009534 |
|            GaussianNB            | 0.8227305926848203 | 0.0014732302562905234 |
+----------------------------------+------------------+----------------------+
```

Fig. 8: Cross-Validation Results

Through these efforts, it was possible to achieve wide-ranging benchmarks thereby providing an overview regarding the performance and steadiness of every algorithm. The cross-validation results indicated that the tree-based classifiers seem to be the most effective. However, we would be employing various classifier types for this project to compare models and do evaluations across them. Five models were identified from the CV in terms of performance ranking for further assessment: top-ranking models were XGBoost, CatBoost, LighGBM from top performers, K-Nearest Neighbors from mid-range performers, and Logistic Regression from low performers.

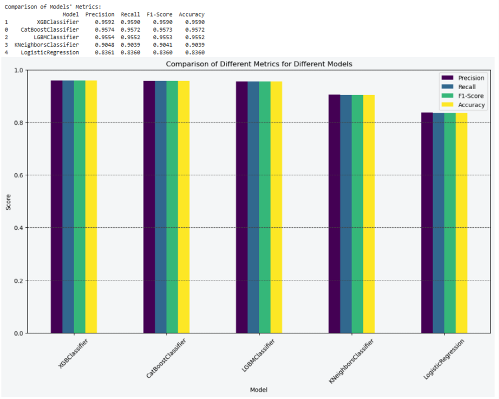
The selected models were then trained on an 80% training split of the dataset and evaluated on the remaining 20% testing set. Then, the models were parameterized and evaluated on the test for their generalization capabilities. In this stage, there was rather little adjustment, with default parameters such as (iterations=100) for the CatBoost, and (n_estimators=100) for the XGBoost and LightGBM. While there was no extreme hyperparameter tuning, for instance, through grid or random search, this decision made a compromise between computational expense and effectiveness.



```
Comparison of Models' Training and Testing Scores:
                   Model  Training Score  Testing Score
1          XGBClassifier        0.969260       0.958962
0     CatBoostClassifier        0.964689       0.957230
2         LGBMClassifier        0.956951       0.955228
3   KNeighborsClassifier        0.931697       0.903950
4     LogisticRegression        0.835464       0.835964
```

Fig. 9: Training vs. Testing Scores for Different Models

The evaluation was done in several ways to be more comprehensive in determining the effectiveness of the models which are the following: accuracy, precision, recall, F1–score, and AUC-ROC. The performance of the tree-based ensemble methods was superior with the XGBoost model recording the highest test accuracy of 95.90% and a subsequent AUC of 99.42% for the model, which was very close to that of CatBoost and LightGBM. These models managed to explain many nonlinear forms and interactions between features better than simpler models such as Logistic Regression which had an accuracy of 83.60%.



```
Comparison of Models' Metrics:
                   Model  Precision  Recall  F1-Score  Accuracy
1          XGBClassifier     0.9592  0.9590    0.9590    0.9590
0     CatBoostClassifier     0.9574  0.9572    0.9573    0.9572
2         LGBMClassifier     0.9554  0.9552    0.9553    0.9552
3   KNeighborsClassifier     0.9048  0.9039    0.9041    0.9039
4     LogisticRegression     0.8361  0.8360    0.8360    0.8360
```

Fig. 10: Comparison of Different Metrics for Different Models

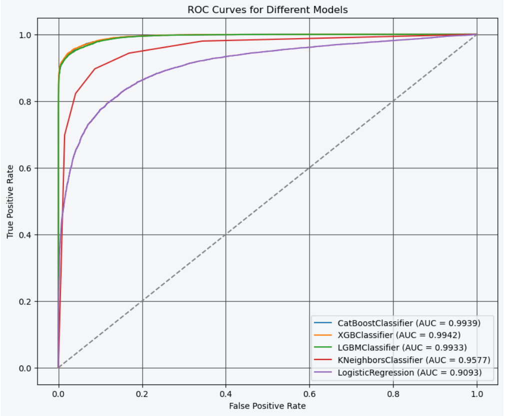


Fig. 11: ROC Curves for Different Models

The confusion matrices verified that ensemble approaches reduced both false positives and false negatives while K-Nearest Neighbors showed signs of overfitting, with larger discrepancies between training and testing performance.

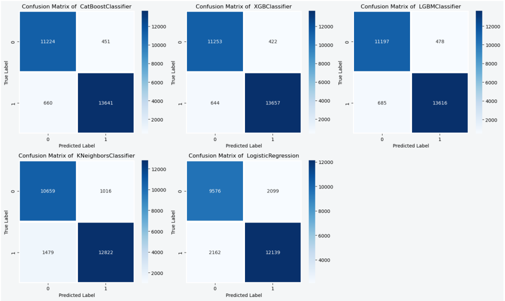


Fig. 12: Confusion Matrices of Different Models

B. *Energy Appliances Analysis*

1) Business Understanding: The aim was to predict appliance energy consumption (Appliances) while appreciating variables that contribute to energy usage including temperature, humidity, time, and day in a month. Since the user is able to make accurate predictions, it helps to aid the energy efficiency measures in smart homes hence the analysis can be relevant to the achievement of sustainable goals.

2) Data Understanding: The dataset constituted 29 variables and contained 19,735 records that related to time and the environment. Initial assessment confirmed that there are no data missing or duplicates which makes the dataset useful. Interrelation analysis showed variation across the indoor temperatures (T1-T9), RH levels (RH_1-RH_9), and selected weather features like windspeed and visibility.

Target Variable: The variable of interest, Appliances, was very skewed and therefore needed to be transformed.

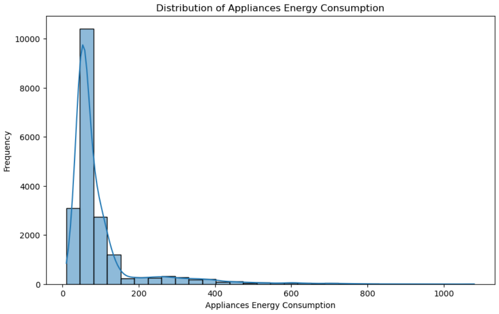


Fig. 13: Distribution of Appliances Energy Consumption

Temporal Trends: The results of the energy consumption analysis were able to derive an hourly trend which revealed that the energy was relatively high during the evening and low at night, however, there was no clear or unambiguous seasonality pattern across the daily, weekly or monthly averages. The lack of strong seasonality or patterns over time series analysis meant that, most probably, other factors, not time, affected energy consumption, thus loading regression or machine learning based models were more appropriate than time series approaches.

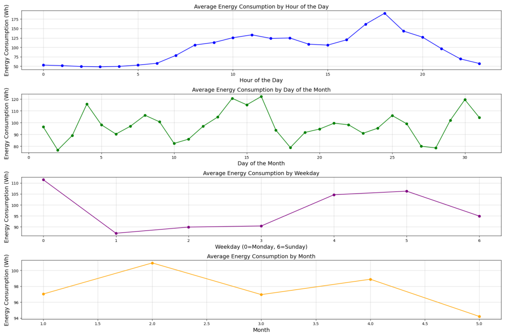


Fig. 14: Temporal Trends

Outliers Analysis: A total of 2,138 values were outliers in the target variable. These outliers should have been included as they seemed to be genuine cases of extreme usage and were necessary to capture the variation. Target variables should not be excluded. Such inclusion was important in ensuring that the model trained and validated had more generalization and prediction power when in high consumption circumstances. Their exclusion would bear the risk of having an 'average' model that loses important details.

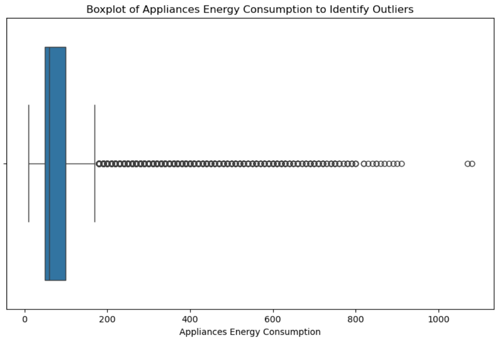


Fig. 15: Boxplot of Appliances Energy Consumption

3) Data Preparation: Preprocessing was critical to ensure data quality and improve model accuracy.

Temporal Features Engineering: The date column was in string format while hour, day, weekday, and month were automatically extracted.

Feature Grouping: Features were organized according to their logic: temperature (T1-T9), humidity (RH_1-RH_9), weather (T_out, RH_out, windspeed, visibility), light usage (lights), and target (Appliances).

Feature Transformation: A log transformation was applied to normalize the right-skewed distribution of the target variable. This step assisted in concentrating the distribution which enhanced linear algorithms through transformation of the target to improve the model fit.

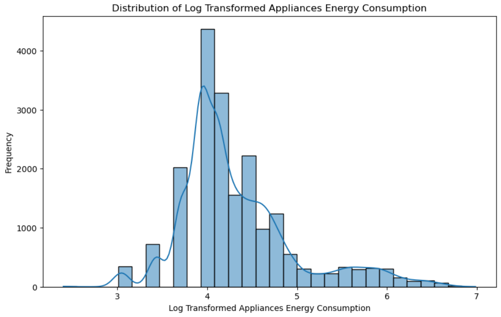


Fig. 16: Distribution of Log Transformed Appliances Energy Consumption

Correlation Analysis: The heatmap revealed some features having strong correlations such as rv1 and rv2, T9 with other temp features. To counter the effect of redundancy rv2, T6, T9, RH_8 were dropped due to having high correlation values. Also, lights feature was also omitted because the majority of the "lights" variable contain zero values indicating a positively skewed distribution.

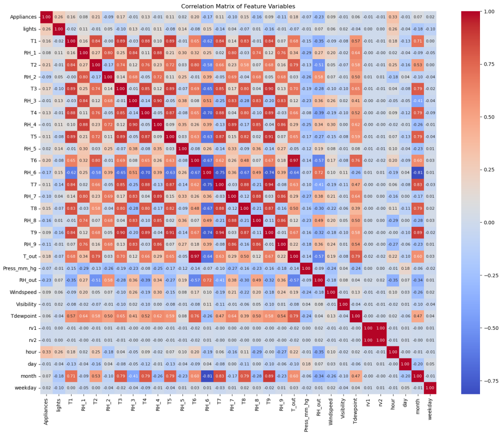


Fig. 17: Correlation Matrix of Feature Variables

Feature Scaling: All numerical features were normalized using StandardScaler to enhance compatibility with algorithms sensitive to scale (e.g.,Ridge Regression, Support Vector Regression).

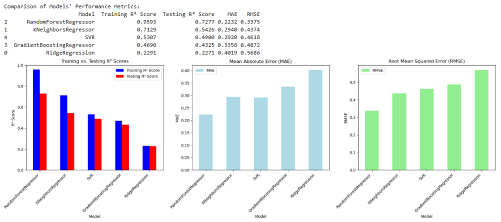
4) Modeling & Evaluation: The evaluations that were carried out on the models were aimed at examining the effectiveness of the techniques that were employed to forecast the appliance energy consumption. The evaluation phase included determining the relevant performance metrics, hyperparameter optimization, model selection, and result analysis in order to come up with explanations. The evaluation process and results are outlined below.

Performance Metrics: The models that were put through testing include Ridge Regression, K-Nearest Neighbors, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Regressor.

The Random Forest Regressor achieved the highest results, with a training R^2 of 0.9593 while the testing R^2 value stood at 0.7277. However, the difference that exists between the training and testing scores suggests overfitting has occurred. This challenge was to some extent resolved through hyperparameter tuning and feature selection, thus improving its performance.

K-Nearest Neighbors Regressor managed to have a lesser overfitting gap, but the accuracy was found to be lower. Good generalization was also observed in Support Vector Regressor and Gradient Boosting Regressor which had a diminutive bias, and low training–testing gaps, but the complexity handling was still not on a par with Random Forest. While Ridge Regression performed woefully low because it is a linear one, and the data set had a lot of nonlinear relationships.

While examining the various evaluation metrics across the models, it was observed that Random Forest had an MAE (Mean Absolute Error) of 0.2232 and an RMSE (Root Mean Squared Error) of 0.3375 implying that the model predictions had a lower dispersion around the actual values. On the other hand, Ridge Regression had a much higher RMSE value which is to say that the performance of the model on the dataset was poor.



```
Comparison of Models' Performance Metrics:
                      Model  Training R² Score  Testing R² Score    MAE   RMSE
2     RandomForestRegressor             0.9593            0.7277 0.2232 0.3375
1       KNeighborsRegressor             0.7129            0.5426 0.2940 0.4374
4                       SVR             0.5307            0.4900 0.2920 0.4618
3 GradientBoostingRegressor             0.4690            0.4325 0.3358 0.4872
0           RidgeRegression             0.2291            0.2271 0.4019 0.5686
```

Fig. 18: Comparison of Models' Performance Metrics

Hyperparameter Tuning: Hyperparameter tuning was emphasized using the Random Forest model for its improvement purposes. The best parameters were found using GridSearchCV as follows: Number of estimators: 200. Maximum depth: None. Minimum samples per split: 2. Minimum samples per leaf: 1. After tuning the parameters, the R^2 score for the performance of the new improved Random Forest model was equal to 0.7423, MAE score of 0.2161, and RMSE score of 0.3283.

Feature Importance: A feature importance analysis was carried out to establish what features were the most salient features that were responsible for the predictions. This step was very important as it explained the contribution of different features to the energy consumption prediction model. "Hour" turned out to be the most significant variable which is consistent with the expectation that energy consumption patterns are dependent on the time of the day.

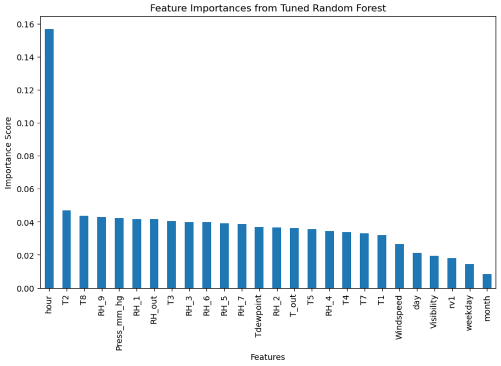
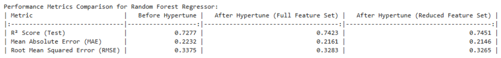


Fig. 19: Feature Importance from Tuned Random Forest

Feature Reduction: In an effort to increase the accuracy of the model, features that had an importance of less than 0.02 were excluded which increased accuracy but didn't reduce performance. This resulted in a testing R^2 value of 0.7451, an MAE of 0.2146, and an RMSE of 0.3265 which were modest improvements from the model that was not tuned. These adjustments highlight the importance of reducing the dimensionality of a model when seeking to enhance model predictive power whilst not losing the ability to explain the model.

Based on the assessment, it could be concluded that the optimized model of Random Forest did not only improve the accuracy of predictions but also provided useful information about feature contributions, thereby proving to be a model that is adequate for the work of predicting the energy consumption of the appliance.



```
Performance Metrics Comparison for Random Forest Regressor:
| Metric                        |   Before Hypertune |   After Hypertune (Full Feature Set) |   After Hypertune (Reduced Feature Set) |
|:------------------------------|-------------------:|-------------------------------------:|----------------------------------------:|
| R² Score (Test)               |             0.7277 |                               0.7423 |                                  0.7451 |
| Mean Absolute Error (MAE)     |             0.2232 |                               0.2161 |                                  0.2146 |
| Root Mean Squared Error (RMSE) |             0.3375 |                               0.3283 |                                  0.3265 |
```

Fig. 20: Performance Metrics Comparison for Random Forest Regressor

C. Twitter Sentiment Analysis

1) Business Understanding: The aim was to separate tweets with information into four different sentiment categories, those being — Positive, Negative, Neutral, and Irrelevant — Keeping in mind, informal language, unbalanced sentiment categories, and noise in the data. The strong preprocessing of the data and selection of the appropriate model helped in achieving the goal of this task which was sentiment classification.

2) Data Understanding: The dataset consisted of approximately 74,681 records, capturing tweet IDs, associated entities, sentiment labels, and tweet content. Sentiments were categorized as Positive, Negative, Neutral, and Irrelevant.

Sentiment Distribution: The bar and the pie charts were employed to depict the distribution of sentiment. It was evident from the graphs that the most ordinary sentiment expressed was at 30.3% Negative, remaining at 27.5% Positive sentiment, 24.7% Neutral sentiment, and 17.5% Irrelevant sentiment.

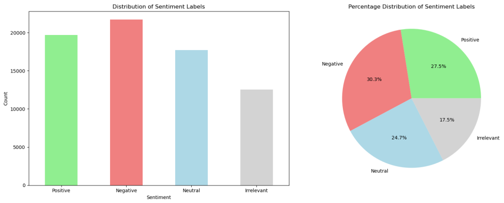


Fig. 21: Distribution of Sentiment Labels

Entity Distribution: This count noted that some of the entities such as ”TomClancysRainbowSix” and ’Verizon” have a huge following. A bar plot was created to show this distribution so as to help in knowing the entities that had most mentions.

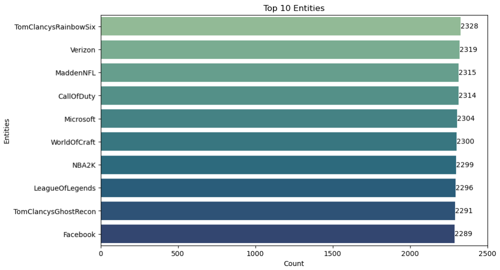


Fig. 22: Entity Distribution

Tweet Length: The fact that tweet lengths varied very little between sentiment categories suggests that length was not a crucial component in sentiment classification. The significance of managing extreme cases for consistency was shown by outliers in every category. For sentiment analysis to be reliable, content was more important than length.

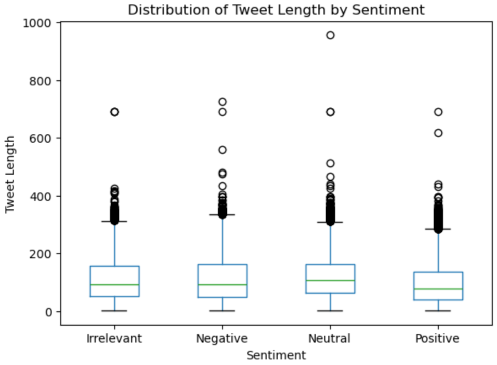


Fig. 23: Distribution of Tweet Length by Sentiment

Word Clouds: Created for each class of sentiment, depicting the key words which were used in the tweets. For example, “game”, “love” and “good” are more likely to be used in Positive tweets whilst “now”, “shit” and “fix” are found in Negative tweets.

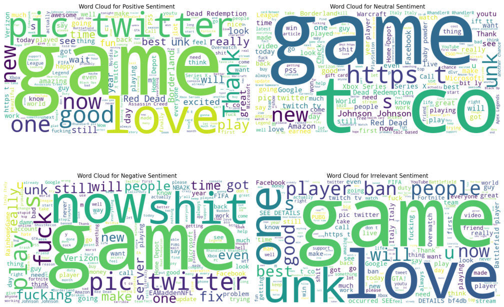


Fig. 24: Word Cloud of Sentiment Labels

3) Data Preparation: Proper preprocessing was necessary to obtain high-quality datasets, as it lowered the computational load and enables the machine learning model to identify only pertinent data.

Handling Missing Values: Any rows that missing Tweet in their fields were deleted because textual data is very important in sentiment analysis.

Duplicate Removal: Duplicate rows were deleted so that data breakdown was minimized plus the amount of data available for analysis was less. This provided the availability of data without any duplication.

Normalization: Using text in lower case was sufficient to standardize the text and reduce the amount of noise.

Cleaning: Eliminating numeric characters, HTML, URLs, punctuation, and emoticons sorts unwanted elements from the tweets.

Stopword Removal: Words that appear to be the most frequent but make insignificant meaningful contribution to the analysis (e.g., ‘is’, the’) were deleted so that the model relies more on meaningful words.

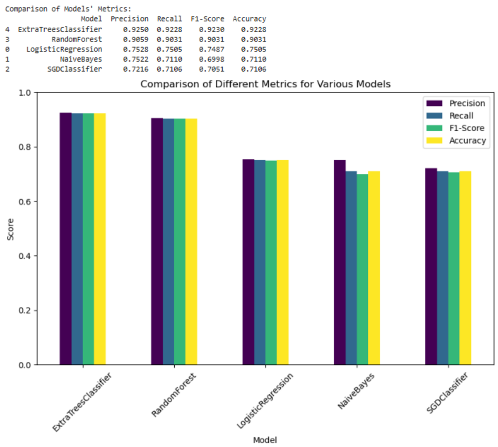
Tokenization and Lemmatization: This involves breaking the tweets into individual words (tokenization) and reducing them to their base or root forms (lemmatization), so that variations of a word, such as 'run' and 'running,' are recognized as the same term to retain their underlying meaning.

Feature Engineering: TF-IDF Vectorization—This phase transformed the text data into numerical representations, assigning higher importance to less frequent but more meaningful keywords, which are critical for efficient sentiment classification.

4) Modeling & Evaluation: The insights gained from the comparison of multiple machine learning models for sentiment classification regarding their effectiveness were tremendous. The main points are aggregated and presented below.

Model Comparison and Overall Performance: The Extra Trees Classifier and the Random Forest Classifier were the best models, performing well in terms of accuracy, precision, recall, and F1 scores (more than 90%). These two models also performed well on the test set which indicates a strong correlation with the training set.

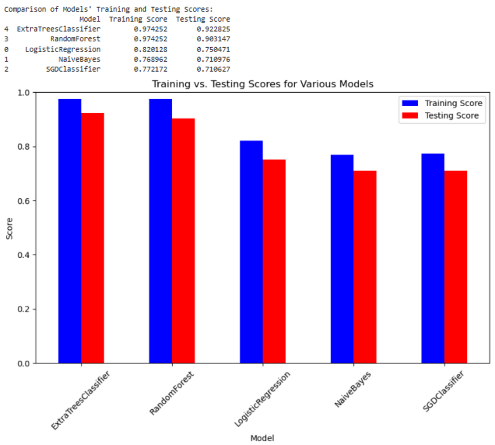
Other models including Logistic Regression and Naïve Bayes performed averagely with accuracy of about 75%. Such models are relatively easier to compute but are weak when it comes to dealing with nonlinear multidimensional relationships. The performance of the SGD Classifier was somewhat close to that of Naïve Bayes, confirming that it is an effective linear classifier, which is simple in design but quite robust.



```
Comparison of Models' Metrics:
                   Model  Precision  Recall  F1-Score  Accuracy
4  ExtraTreesClassifier     0.9250  0.9228    0.9230    0.9228
3          RandomForest     0.9059  0.9031    0.9031    0.9031
0    LogisticRegression     0.7528  0.7505    0.7487    0.7505
1            NaiveBayes     0.7522  0.7110    0.6998    0.7110
2         SGDClassifier     0.7216  0.7106    0.7051    0.7106
```

Fig. 25: Comparison of Different Metrics for Various Models

Overfitting Analysis: Extra Trees Classifier demonstrated the least overfitting and best overall performance, making it the most suitable model for this dataset. Random Forest Classifier showed signs of mild overfitting and overfitting is evident in Logistic Regression, Naïve Bayes, and SGD Classifier due to their larger training-testing score gaps.



```
Comparison of Models' Training and Testing Scores:
                   Model  Training Score  Testing Score
4  ExtraTreesClassifier        0.974252       0.922825
3          RandomForest        0.974252       0.903147
0    LogisticRegression        0.820128       0.750471
1            NaiveBayes        0.768962       0.710976
2         SGDClassifier        0.772172       0.710627
```

Fig. 26: Training vs. Testing Scores for Various Models

Confusion Matrices Insights: The models had difficulties differentiating between neutral or irrelevance sentiments because of the semantic similarities, while Extra Trees and Random Forest in most of the cases correctly classified the highest number of items across the sentiment classes.

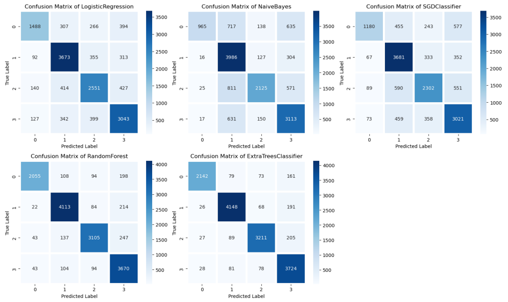


Fig. 27: Comparison of Models' Confusion Matrix

Sentiment Distribution and Prediction Alignment: The comparison of the actual sentimental values with the estimated sentiments revealed that the Extra Trees Classifier was almost equally near the true distributions, exhibiting its tendencies to adequately control the class of predictions. SGD Classifier, Logistic Regression, and Naïve Bayes were over-predicting some specific sentiments and in particular, the class termed as irrelevant.

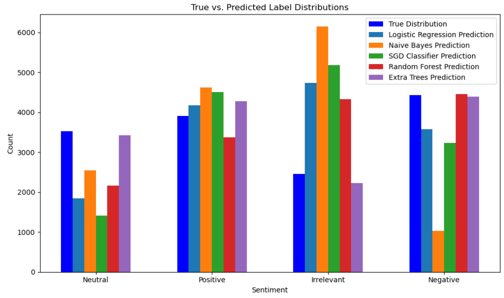


Fig. 28: True vs. Predicted Label Distributions

IV. Conclusions and future work

In this work, three datasets—Airline Passenger Satisfaction, Energy Appliance Consumption, and Twitter Sentiment Analysis—were used to implement machine learning models with the aim of finding the drivers of the outcomes and the performance of the predictions. This trend continued with other datasets as well when ensemble models such as XGBoost, Random Forest, and Extra Trees were used, which only proved better results, stressing their applicability across a wide range of data structures.

As for the Airline Passenger dataset, it was found that service aspects such as Entertainment and Seat Comfort contributed dominantly toward achieving satisfaction hence providing airlines with great help in formulating strategies to meet customer satisfaction. The Energy dataset revealed the "Hour of the Day" as the most important factor influencing energy consumption while demand side variables such as weather, were less important. The solutions in the Twitter Sentiment dataset, such as an ensemble of TF-IDF vectorization and other techniques, were effective in the analysis to manage issues arising from the use of informal language and the imbalance in the sentiment classes.

However there were some constraints that the study had to grapple with. The existence of subjectivism ratings in the airline data set was a source of perhaps inconsistencies, while the short-term nature of the energy data set limited the study of seasonal patterns. Avoiding the more complicated algorithms was done because of the computing resources limitation. Also, the amount of Twitter data, which was unstructured language, required a considerable amount of preprocessing to eliminate any information which might have led to the loss of critical context.

Future directions should address these constraints to enhance the analysis further. For the Energy dataset, the inclusion of additional variables such as long-term trends or holiday indicators could enhance seasonality and thus boost the forecasts. Concerning the Twitter dataset, Twitter word embeddings and advanced natural language processing approaches based on transformers, for example, BERT, may replace the Twitter word embeddings to capture the underlying Twitter nuances better. Finally, incorporating real-time feedback in the airline dataset could enhance satisfaction prediction by making it more applicable. The accuracy and level of generalization could be improved if more effort was put in altering the hyperparameters and tuning deep learning models.

To sum up, this study demonstrates how crucial is the ensemble methods, data preprocessing and feature engineering techniques when working with different datasets. The outcomes help practitioners with ways of strategies formulation which enhance customer satisfaction, improve energy efficiency, and reliable sentiment analysis, while causing for more investigations directed at overcoming the demonstrated limitations and expanding the depth of analysis.

References

[1] L. Hibović, S. Smajić and E. Yaman, "Predicting Satisfaction of Airline Passengers Using Classification," 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2022, pp. 939-946, doi: 10.1109/ISMSIT56059.2022.9932850. [Online]. Available: https://ieeexplore.ieee.org/document/9932850

[2] Z. Shu, "Analysis of Flight Delay and Cancellation Prediction Based on Machine Learning Models," 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2021, pp. 260-267, doi: 10.1109/MLBDBI54094.2021.00056. [Online]. Available: https://ieeexplore.ieee.org/document/9731090

[3] C. Tan, "Bidirectional LSTM Model in Predicting Satisfaction Level of Passengers on Airline Service," 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE), Hangzhou, China, 2021, pp. 525-531, doi: 10.1109/ICAICE54393.2021.00107. [Online]. Available: https://ieeexplore.ieee.org/document/9797466

[4] J. Zhao, "Research on Evaluation of Airline Passenger Satisfaction Based on Fuzzy Comprehensive Evaluation," 2023 5th International Conference on Applied Machine Learning (ICAML), Dalian, China, 2023, pp. 113-117, doi: 10.1109/ICAML60083.2023.00030. [Online]. Available: https://ieeexplore.ieee.org/document/10457501

[5] Kadir Amasyali, Nora M. El-Gohary, "A review of data-driven building energy consumption prediction studies", Renewable and Sustainable Energy Reviews, Volume 81, Part 1, 2018, Pages 1192-1205, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2017.04.095. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032117306093

[6] Tanveer Ahmad, Huanxin Chen, "A review on machine learning forecasting growth trends and their real-time applications in different energy systems, Sustainable Cities and Society, Volume 54, 2020, 102010, ISSN 2210-6707, https://doi.org/10.1016/j.scs.2019.102010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210670719335516

[7] L. Liu, F. H. Juwono, W. K. Wong and H. Liu, "Building Energy Consumption Prediction: A Machine Learning Approach with Feature Selection," 2024 10th International Conference on Smart Computing and Communication (ICSCC), Bali, Indonesia, 2024, pp. 159-164, doi: 10.1109/ICSCC62041.2024.10690314. [Online]. Available: https://ieeexplore.ieee.org/document/10690314

[8] S. K. Mohapatra, S. Mishra and H. K. Tripathy, "Energy Consumption Prediction in Electrical Appliances of Commercial Buildings Using LSTM-GRU Model," 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), Bhubaneswar, India, 2022, pp. 1-5, doi: 10.1109/ASSIC55218.2022.10088334. [Online]. Available: https://ieeexplore.ieee.org/document/10088334

[9] Z. Jianqiang, "Pre-processing Boosting Twitter Sentiment Analysis?," 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), Chengdu, China, 2015, pp. 748-753, doi: 10.1109/SmartCity.2015.158. [Online]. Available: https://ieeexplore.ieee.org/document/7463812

[10] A. Sarlan, C. Nadam and S. Basri, "Twitter sentiment analysis," Proceedings of the 6th International Conference on Information Technology and Multimedia, Putrajaya, Malaysia, 2014, pp. 212-216, doi: 10.1109/ICIMU.2014.7066632. [Online]. Available: https://ieeexplore.ieee.org/document/7066632

[11] M. Khurana, A. Gulati and S. Singh, "Sentiment Analysis Framework of Twitter Data Using Classification," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, India, 2018, pp. 459-464, doi: 10.1109/PDGC.2018.8745748. [Online]. Available: https://ieeexplore.ieee.org/document/8745748

[12] L. Mandloi and R. Patel, "Twitter Sentiments Analysis Using Machine Learninig Methods," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9154183. [Online]. Available: https://ieeexplore.ieee.org/document/9154183