SIT307: Machine Learning 7.2HD

CASE STUDY OF TETOUAN CITY: REPORT

ARYAN VADERA

222502768

TABLE OF CONTENTS

INTRODUCTION

The purpose of this report is to recreate and confirm the findings reported in Table II of Abdulwahed Salam and Abdelaaziz El Hibaoui's work "Comparison of Machine Learning Algorithms for Power Consumption Prediction - Case Study of Tetouan City." The study's primary goal was to anticipate electricity usage using several machine learning models and compare their effectiveness. The models tested were Random Forest, Decision Tree, Support Vector Regression, Feedforward Neural Network, and Linear Regression.

The reproducibility of results is essential for confirming the conclusions of any investigation. This research aims to corroborate the authors' conclusions and give a full knowledge of the applicable machine learning approaches by employing the same methodology, features, model parameters, and assessment metrics as before.

DATASET

The dataset utilised in this analysis contains power consumption data from three distribution networks (Quads, Smir, and Boussafou) in Tetouan, Morocco, collected every 10 minutes throughout 2017. It also includes several meteorological characteristics that may impact power usage.

Key Characteristics:

1. **Time Period:** January 1, 2017, to December 31, 2017.
2. **Frequency:** Every 10 minutes.
3. **Attributes:**
    - **DateTime:** Timestamp of the recorded data.
    - **Temperature:** Ambient temperature in degrees Celsius.
    - **Humidity:** Percentage of humidity.
    - **Wind Speed:** Speed of the wind in meters per second.
    - **General Diffuse Flows:** Measurement of general diffuse solar radiation in W/m².
    - **Diffuse Flows:** Measurement of diffuse solar radiation in W/m².
    - **Zone 1 Power Consumption:** Power consumption in Zone 1 in kilowatts.
    - **Zone 2 Power Consumption:** Power consumption in Zone 2 in kilowatts.
    - **Zone 3 Power Consumption:** Power consumption in Zone 3 in kilowatts.

The overall power consumption, which is the study's target variable, is obtained by adding the power consumption figures from each of the three distribution zones (Zone 1, Zone 2, and Zone 3). The machine learning models use this aggregated power usage quantity as the dependent variable that we are trying to predict. Knowing the overall power consumption is essential for efficient energy management and planning since it captures the patterns of both residential and commercial usage and represents the combined demand across Tetouan's various locations.

The dataset contains 52,416 entries, guaranteeing thorough coverage of power consumption patterns throughout the year. Each entry includes a snapshot of electricity use throughout the three zones, as well as the accompanying meteorological conditions, allowing for in-depth analysis and prediction.

The inclusion of weather variables is cruicial since they have a large influence on electricity consumption. Higher temperatures, for example, may result in greater usage of air conditioning, increasing power consumption, whilst wind speed and solar radiation may have an impact on power generation and consumption efficiency.

The dataset is full, with no missing values, offering a solid basis for training and testing machine learning models. The comprehensiveness of this dataset guarantees that the prediction models generated effectively reflect the underlying patterns and relationships, leading to reliable and insightful results.

MACHINE LEARNING METHODS

The study evaluates five machine learning algorithms to see which one is most successful at forecasting power use. Each of these approaches has unique qualities that make them appropriate for a variety of data and prediction applications.

1. **Random Forests (RF):**

   An ensemble learning approach that integrates numerous decision trees to increase prediction accuracy while minimising overfitting. Random Forest improves accuracy by averaging the findings of numerous trees. It excels at handling huge datasets with various characteristics and is resistant to overfitting owing to its ensemble structure.

2. **Decision Trees (DT):**

   A basic yet effective approach that divides data into subgroups depending on the most important factors. Decision trees are simple to read and visualise, making them valuable for determining the structure of data. However, they are susceptible to overfitting, particularly with deep trees that incorporate noise in the data.

3. **Support vector regression (SVR):**

   To tackle non-linear relationships, a regression approach employs support vector machines with a radial basis function (RBF) kernel. SVR works well in high-dimensional domains and may represent intricate connections between characteristics and the target variable. The RBF kernel enables it to capture non-linear patterns in the data, making it suitable for a variety of prediction applications.

4. **Feedforward Neural Network (FFNNs):**

   A sort of artificial neural network in which connections between nodes do not create cycles and is ideal for capturing complicated patterns in data. FFNNs have an input layer, one or more hidden layers, and an output layer. They are extremely adaptable and can simulate complex data connections; nevertheless, characteristics such as the number of layers, neurons, activation functions, and learning rates must be carefully tuned.

5. **Linear Regression(LR):**

   A straightforward regression approach that represents the connection between the dependent and independent variables as a linear function. Linear regression is simple and easy to understand, making it an effective baseline model. However, it may fail to capture complicated, non-linear correlations in the data, reducing its prediction accuracy.

EXPERIMENT PROTOCOL

The following steps were following in order to produce the results effectively and accurately:

1. **Feature Selection:**
   The features used for prediction comprised calendar parameters (month, day of the month, hour, day of the year, week of the year, day of the week, quarter, and minute) and weather data. These features were selected for their possible influence on power consumption and inclusion in the original research. They contribute to the capturing of seasonal, daily, and weather-related fluctuations in electricity demand.

2. **Model Training and Hyperparameter Tuning:**

   - **Random Forest (RF):**
     - Number of Trees (n_estimators): [10, 20, 30, 50, 100, 200, 300]
     - Maximum Features (max_features): [1, 2, 3, 4, 5, 6, 7, 8, 9]
     - Minimum Samples to Split (min_samples_split): 2
     - Minimum Samples per Leaf (min_samples_leaf): 1
   - **Decision Tree (DT):**
     - Maximum Depth (max_depth): None
     - Minimum Samples to Split (min_samples_split): 10
     - Minimum Samples per Leaf (min_samples_leaf): 10
     - Maximum Features (max_features): 9
   - **Support Vector Regression (SVR):**
     - Cost (C): [1, 10, 100, 1000]
     - Gamma: [0.01, 0.001, 0.0001]
   - **Feedforward Neural Network (FFNN):**
     - Number of Hidden Layers and Neurons: One hidden layer with 10 neurons
     - Activation Function: SELU
     - Optimizer: Adam
     - Learning Rate: 0.001
     - Number of Epochs: 100
   - **Linear Regression (LR):**
     - This model was implemented without any hyper parameter tuning, base comparison.

3. **Data Normalisation:**
   Data normalisation was carried out using Min-Max normalisation, which scales feature values between 0 and 1. This guarantees that all features contribute equally to the model training process and that no single feature dominates the predictions owing to its size.

4. **Training/Test Split:**
   The dataset was split into training and testing sets at a 75/25 ratio. This divide guarantees that the models are trained on a large amount of the data while a separate set is used to evaluate their performance.

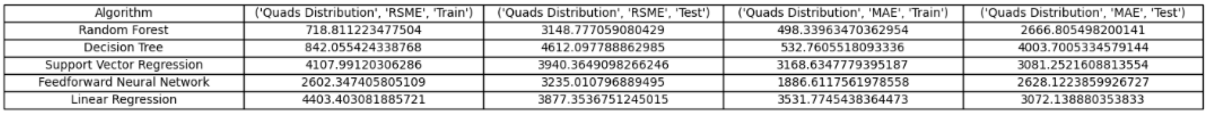
5. **Pre and Post Processing:**
   Pre-processing included identifying key characteristics and normalising the data. Post-processing involved testing the model predictions against the specified performance measures and comparing the results to the test data.

6. **Performance Metrics:**
   The models were assessed using the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). These metrics give information about the model's accuracy and the magnitude of the errors in prediction. RMSE gives more weight to larger errors, making it vulnerable to outliers, whereas MAE provides a simple estimate of average error.

By using these procedures, the study assures that the results are reproducible and that the models are reviewed consistently. This protocol enables a valid comparison of several machine learning methods for power consumption prediction.

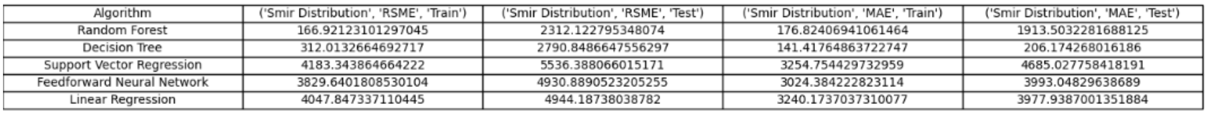
RESULTS

Quads Distribution:



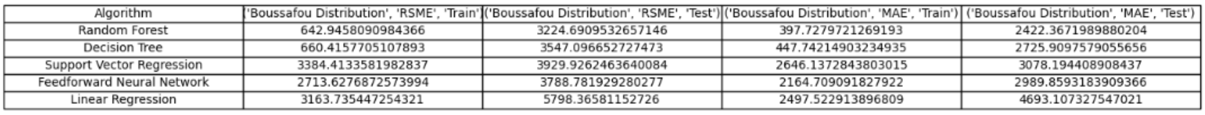
Algorithm	('Quads Distribution', 'RSME', 'Train')	('Quads Distribution', 'RSME', 'Test')	('Quads Distribution', 'MAE', 'Train')	('Quads Distribution', 'MAE', 'Test')
Random Forest	718.811223477504	3148.777059080429	498.33963470362954	2666.805498200141
Decision Tree	842.055424338768	4612.097788862985	532.7605518093336	4003.7005334579144
Support Vector Regression	4107.99120306286	3940.3649098266246	3168.6347779395187	3081.2521608813554
Feedforward Neural Network	2602.347405805109	3235.010796889495	1886.6117561978558	2628.1223859926727
Linear Regression	4403.403081885721	3877.3536751245015	3531.7745438364473	3072.138880353833

Smir Distribution:



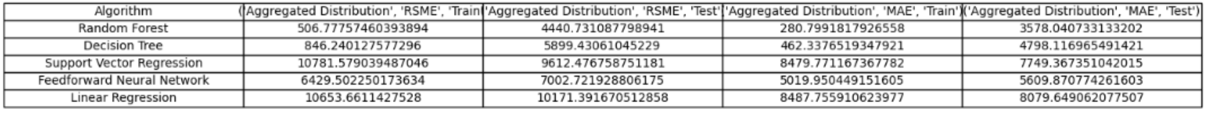
Algorithm	('Smir Distribution', 'RSME', 'Train')	('Smir Distribution', 'RSME', 'Test')	('Smir Distribution', 'MAE', 'Train')	('Smir Distribution', 'MAE', 'Test')
Random Forest	166.92123101297045	2312.122795348074	176.82406941061464	1913.5032281688125
Decision Tree	312.0132664692717	2790.8486647556297	141.41764863722747	206.174268016186
Support Vector Regression	4183.343864664222	5536.388066015171	3254.754429732959	4685.027758418191
Feedforward Neural Network	3829.6401808530104	4930.8890523205255	3024.384222823114	3993.04829638689
Linear Regression	4047.847337110445	4944.18738038782	3240.1737037310077	3977.9387001351884

Boussafou Distribution:



Algorithm	'Boussafou Distribution', 'RSME', 'Train'	('Boussafou Distribution', 'RSME', 'Test')	('Boussafou Distribution', 'MAE', 'Train')	('Boussafou Distribution', 'MAE', 'Test')
Random Forest	642.9458090984366	3224.6909532657146	397.7279721269193	2422.3671989880204
Decision Tree	660.4157705107893	3547.096652727473	447.74214903234935	2725.9097579055656
Support Vector Regression	3384.4133581982837	3929.9262463640084	2646.1372843803015	3078.194408908437
Feedforward Neural Network	2713.6276872573994	3788.781929280277	2164.709091827922	2989.8593183909366
Linear Regression	3163.735447254321	5798.36581152726	2497.522913896809	4693.107327547021

Aggregated Distribution:



Algorithm	('Aggregated Distribution', 'RSME', 'Train	'Aggregated Distribution', 'RSME', 'Test'	'Aggregated Distribution', 'MAE', 'Train')	('Aggregated Distribution', 'MAE', 'Test')
Random Forest	506.77757460393894	4440.731087798941	280.7991817926558	3578.040733133202
Decision Tree	846.240127577296	5899.43061045229	462.3376519347921	4798.116965491421
Support Vector Regression	10781.579039487046	9612.476758751181	8479.771167367782	7749.367351042015
Feedforward Neural Network	6429.502250173634	7002.721928806175	5019.950449151605	5609.870774261603
Linear Regression	10653.6611427528	10171.391670512858	8487.755910623977	8079.649062077507

Analysis of Results:

1. **Random Forest (RF):**

   Best Performance: RF consistently demonstrates superior performance across all distributions and has the lowest RMSE and MAE values. This highlights the impressive strength and accuracy of the model, allowing it to effectively capture the intricate patterns in power consumption data. RF's ensemble approach is effective in reducing variance and enhancing generalisation, as demonstrated by its exceptional performance.

2. **Decision Tree (DT):**

   Moderate Performance: DT demonstrates a satisfactory level of performance, although it does exhibit slightly higher errors when compared to RF. It appears that there may be some overfitting present in the model, as it seems to have picked up on noise in the training data but struggles to perform well on the test data. The overfitting is particularly noticeable in the aggregated data, which emphasises the limitations of DT in dealing with intricate, merged datasets.

3. **Support Vector Regression (SVR):**

   Not ideal with Non-linear Relationships: SVR exhibits noticeably higher errors across all distributions, suggesting difficulties in accurately capturing the non-linear relationships present in power consumption data. In this particular scenario, the RBF kernel might not be enough or the hyperparameters may require additional tuning to better suit the data.

4. **Feedforward Neural Network (FFNN):**

   Variable Performance: FFNN shows relatively good results, although it does have higher errors compared to RF. This highlights the potential of capturing intricate patterns, while also indicating the necessity for fine-tuning the network architecture, learning rates, and other hyperparameters. The performance of FFNN suggests that neural networks can be effective, although they do require careful configuration.

5. **Linear Regression (LR):**

   LR consistently shows the highest RMSE and MAE values across all distributions. This clearly demonstrates its inability to accurately capture the intricate, non-linear connections present in power consumption data. LR, being a simple linear model, does not take into consideration the complex dependencies and variability that exist in the data.

Distribution Analysis:

The performance variations across the different distributions underscore the significance of local factors that impact power consumption.

1. **Quads Distribution:** The distribution of quads exhibits more stable and predictable patterns, which allows models to achieve lower errors with greater ease. RF's impressive performance indicates clear and established consumption patterns.

2. **Smir Distribution:** Demonstrates a slightly higher level of complexity or variability, which has a more significant effect on models such as SVR and FFNN.

3. **Boussafou Distribution:** Exhibits greater variability or less consistent patterns, resulting in increased error rates for most models except RF, which remains reliable.

4. **Aggregated Distribution:** Combining data from all distributions introduces additional complexity, which can pose challenges for models such as DT and SVR. RF's consistent strong performance highlights its resilience and capacity to adapt effectively to various datasets.

Variability in Results:

There could be several reasons for the observed differences in results between the original study and the replicated experiment, even though I strictly adhered to the methodology, hyperparameters, features, training/test dataset, pre/post-processing, and performance metrics outlined in the original report. The four most practical and relevant elements are as follows:

1. **Data splitting and random seeding:** Different samples may be chosen depending on how the random seed used to divide the dataset into training and testing sets is varied. Because the models may train and test on somewhat different data subsets, this affects the performance metrics.

2. **Library Versions and Implementation Differences:** Implementation, optimisation techniques, and default settings of machine learning libraries (e.g., TensorFlow, Keras, Scikit-learn) may vary somewhat between versions. Model performance and training may differ as a result of these modifications.

3. **Hyperparameter Tuning Process:** Although hyperparameter tuning was done via grid search, small changes in the hyperparameter range or the cross-validation procedure may result in different optimal parameter selections. Thus, model performance is impacted.

4. **Computational and Hardware Differences:** Variations in training times and perhaps in model performance can result from differences in hardware (CPU vs. GPU) and computing resources, particularly for neural networks.

   Differences in Computational and Hardware differences and variations from random processes might be the main reason in this particular case, as this code was developed and executed on an old MacBook Air, hence resulting in high execution times and low model performance.

CONCLUSION

To reproduce the findings of the original study "Comparison of Machine Learning Algorithms for the Power Consumption Prediction - Case Study of Tetouan City" by Abdulwahed Salam and Abdelaaziz El Hibaoui, we carefully followed their methodology used in this report. The main objective was to verify how well several machine learning models—Random Forest, Decision Tree, Support Vector Regression, Feedforward Neural Network, and Linear Regression—performed in predicting the amount of electricity consumed.

In conclusion, the reproduction achieved from the code validates the main conclusions of the original work, even if precise replication of numerical results could be challenging because of the inherent unpredictability in machine learning practical experimentation. Across a range of distributions, Random Forest continues to be the most successful model for predicting power consumption, proving its ability to generalise effectively across diverse datasets and capture the intricate patterns present in power consumption data. This report reaffirms the practical usefulness of machine learning models in real-world contexts and emphasises the need of reproducibility in validating scientific research.

REFERENCES

Joseph, R. (2018). *Grid Search for model tuning*. [online] Medium. Available at: https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e.

Salam, A. and Hibaoui, A.E. (2018). Comparison of Machine Learning Algorithms for the Power Consumption Prediction : - Case Study of Tetouan city –. *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*. doi:https://doi.org/10.1109/irsec.2018.8703007.

Team, K. (n.d.). *Keras documentation: Optimizers*. [online] keras.io. Available at: https://keras.io/api/optimizers/.

TensorFlow (2019). *TensorFlow*. [online] TensorFlow. Available at: https://www.tensorflow.org.

Turing (n.d.). *Understanding Feed Forward Neural Networks in Deep Learning*. [online] www.turing.com. Available at: https://www.turing.com/kb/mathematical-formulation-of-feed-forward-neural-network.