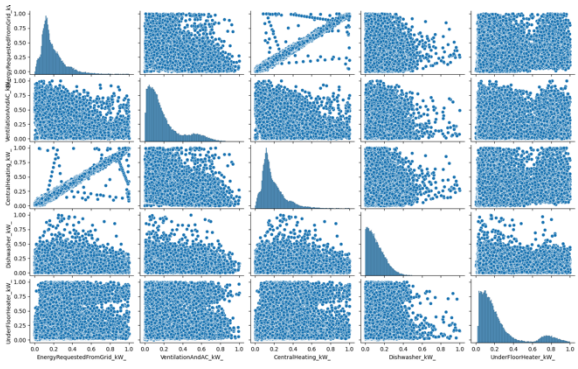
Matthew Jones (Registration Number: 200326702)
Dr Morgan Jones & Dr Giuliano Punzo
ACS341
Sunday 21st April 2024

ACS341 Assignment: Energy Consumption Project

This assignment predicts energy requested from the Grid at a conventional household using two different methods: linear regression using the Ordinary Least Squares (OLS) formula and an Artificial Neural Network. On initial inspection of the dataset, Central Heating kW is almost directly correlated with the value we are ultimately trying to predict, creating a point of confusion about this dataset.



Task 1 – Data Cleaning

For the purposes of data cleaning, I attempted to follow the conventional pre-processing workflow as detailed in the Week 6 lecture on this topic, however I modified the order of processes to fit this dataset better. I started by studying the problem, and removing any columns that were irrelevant to the output, including Radon Level, Wind Bearing and Precipitation Probability.

The dataset was scanned for any values reported as Infinite (Inf) or Not a Number (NaN) and rows containing them were dropped. This was done as the missing values are unlikely to represent any important data that we could not gather from the other data points.

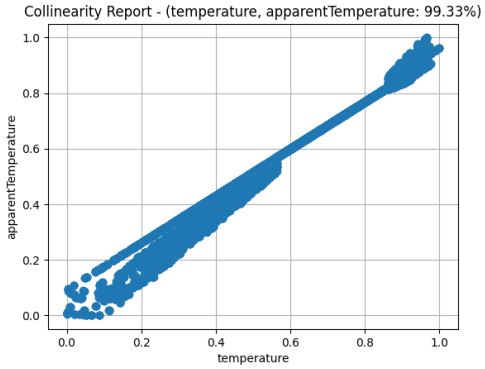
This dataset contained categorical data about the current weather state. One-Hot encoding was chosen to make this numerical, as Integer Encoding assigns "weights" to each different category within the column, which I felt could not accurately represent the data provided. This replaces the categorical column with a series of columns representing each category, containing a Boolean value to state whether that category is present on that row.

Outlier detection was done using the Z-Score method where the number of standard deviations away from the mean of each datapoint is analysed to determine if it is valid data. A Z-Score threshold of 3 was used to decide whether to maintain a data point. For this the workflow was: analyse all the

data points, declare all outliers as NaN values, then, interpolate the missing values based on the other rows. This produces a good approximation for the data without outliers, however it does introduce some additional unpredictable noise into the data.

Min-Max scaling was used to scale each column of data into the range of [0, 1], which ensures that none of the features dominate the model during training. Min-Max was chosen over standardisation as a lot of this data did not fit to the normal distribution, and therefore would not have been scaled well.

Each feature column was then cross compared to check for collinearity. This was done by calculating the Pearson's Correlation Coefficient between the two columns[1], checking to see if it was greater than 0.9 and if so, the second column in the comparison was dropped.



After checking the data for collinearity, the weather-based columns were reduced using Principal Component Analysis (PCA) [2]as initial testing showed that weather data was not very relevant to the final system based on t-statistics of each feature. The 16 weather columns that remain after data processing are reduced to 5 principal components to maintain the data while reducing the number of dimensions being used in the following models.

Task 2 – Linear Regression

For the linear regression, the data is split into 80% training data, and 20% testing data. This is then passed into the statsmodels OLS regressor which generates the following linear regression model.

[1] (Rosidi, 2023)

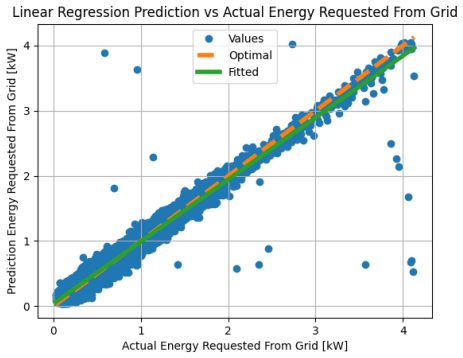
[2] (Galarnyk, 2024)

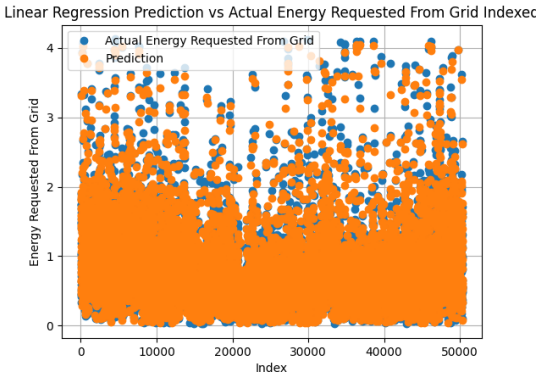
OLS Regression Results

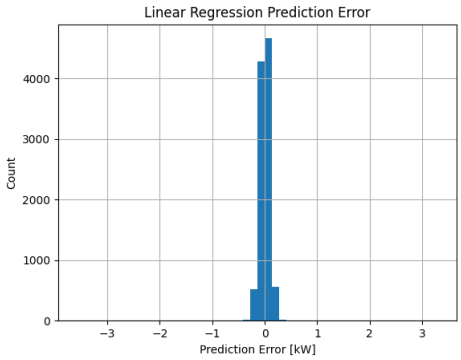
Dep. Variable:	EnergyRequestedFromGrid_kW_	R-squared:	0.960
Model:	OLS	Adj. R-squared:	0.960
Method:	Least Squares	F-statistic:	4.418e+04
Date:	Wed, 24 Apr 2024	Prob (F-statistic):	0.00
Time:	12:02:06	Log-Likelihood:	86046.
No. Observations:	40281	AIC:	-1.720e+05
Df Residuals:	40258	BIC:	-1.718e+05
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0011	0.001	-1.478	0.139	-0.002	0.000
VentilationAndAC_kW_	-0.0036	0.001	-2.624	0.009	-0.006	-0.001
CentralHeating_kW_	0.9751	0.001	760.998	0.000	0.973	0.978
Dishwasher_kW_	0.0044	0.002	2.862	0.004	0.001	0.007
UnderFloorHeater_kW_	0.0055	0.001	7.946	0.000	0.004	0.007
ElectricRadiator_kW_	0.0061	0.001	8.609	0.000	0.005	0.008
HomeOffice_kW_	0.0040	0.001	3.864	0.000	0.002	0.006
Fridge_kW_	0.0016	0.001	2.268	0.023	0.000	0.003
CellarSumpPump_kW_	0.0014	0.001	2.014	0.044	3.68e-05	0.003
GarageDoor_kW_	0.0012	0.002	0.521	0.602	-0.003	0.006
Oven_kW_	0.0003	0.001	0.449	0.653	-0.001	0.002
CeramicHob_kW_	0.0011	0.001	1.206	0.228	-0.001	0.003
Kettle_kW_	0.0003	0.001	0.458	0.647	-0.001	0.002
MechanicalLoftVentilator_kW_	0.0009	0.001	0.783	0.433	-0.001	0.003
ShowerPump_kW_	8.964e-05	0.001	0.071	0.944	-0.002	0.003
Microwave_kW_	0.0003	0.001	0.319	0.750	-0.002	0.002
LivingRoom_kW_	0.0029	0.001	4.385	0.000	0.002	0.004
Solar_kW_	-0.0003	0.001	-0.265	0.791	-0.002	0.002
Weather_PC1	0.0001	0.000	0.566	0.571	-0.000	0.001
Weather_PC2	-0.0011	0.000	-3.069	0.002	-0.002	-0.000
Weather_PC3	-0.0002	0.000	-0.332	0.740	-0.001	0.001
Weather_PC4	-0.0012	0.001	-2.194	0.028	-0.002	-0.000
Weather_PC5	0.0004	0.001	0.741	0.458	-0.001	0.002

Omnibus:	41087.760	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	130791683.611
Skew:	3.791	Prob(JB):	0.00
Kurtosis:	282.052	Cond. No.	22.8







As can be seen from this table most of the solution to the problem is determined by Central Heating kW as shown by the coefficients. This is also shown by the t-statistic being far greater than one with a probability of less than 0.001 (less than 0.05 being optimal).

We can also see that this linear model is a good predictor for the output based on the R^2 statistic of 0.960 (close to 1 is optimal) and the F statistic of 4.410E+04 (greater is better) with a probability of less than 0.01 (less than 0.05 being optimal) as well as the following statistics.

Statistic	Value
Mean Absolute Error (Lower Better)	0.06988
Mean Squared Error (Lower Better)	0.01622
Pearson Correlation between Test Labels and OLS Prediction (Closer to 1 Better)	0.97636

Task 3 – Second Model (Artificial Neural Network)

For this part of the assignment, I opted to use TensorFlow to build an Artificial Neural Network. The model was made up of 3 layers, all of which are densely connected (every node is

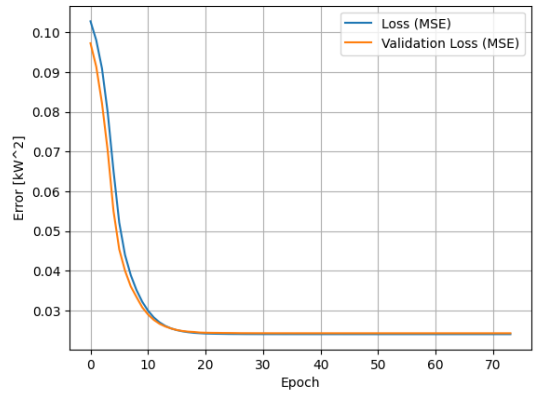
connected to every other node). The units of the first node are automatically tuned by Keras Tuner which is a part of TensorFlow for automatically training the hyperparameters of a model before training. The second layer has 64 units, and the final has 1, to correlate with the number of outputs.

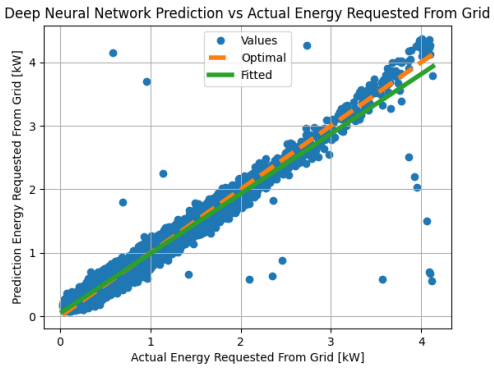
The optimiser of the model is a Stochastic Gradient Descent[3], also using Keras Tuner to tune its learning rate, which uses the Gradient Descent algorithm to minimise loss. For this model, the loss function is Mean Squared Error.

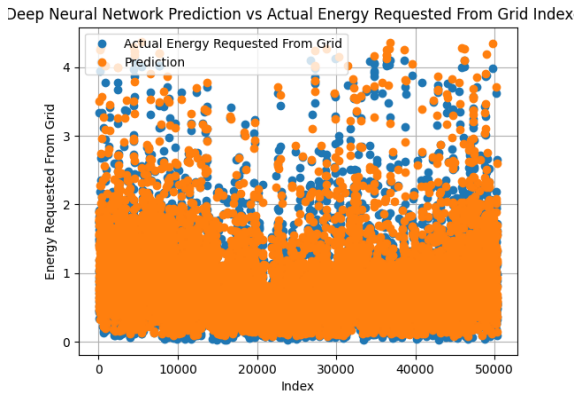
I used several techniques in this process to try and mitigate overfitting, which have worked with varying degrees of success. One technique I employed was Early-Stopping where the model checks as it is training whether the validation loss has stagnated and will cancel the rest of training and restore the best weights. I also limited the Epochs based on the loss graph as well as automatic reduction of learning rate at plateau of validation loss to improve the tuning into the global minimum.

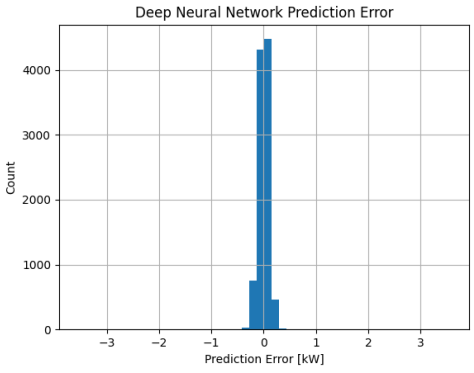
For training, the data was initially split in the same manner as the linear regressor, however the training data was then further split into a ratio of 80% training data and 20% validation data for checking the performance of the model during training.

The following graphs and statistics show the performance of the model.









[3] (TensorFlow, 2024)

Statistic	Value
Mean Absolute Error (Lower Better)	0.07643
Mean Squared Error (Lower Better)	0.01793
Pearson Correlation between Test Labels and OLS Prediction (Closer to 1 Better)	0.97386

From this, we can see that the model performs quite well at predicting the Energy Requested, until about 3kW where it starts to drift to a higher value than reality (in the context of this project, this minor deviation is not too much of a problem, at this level, as it will inform the national grid to increase energy production).

In comparison with the OLS regression, we can see that this model performs worse from the Mean Absolute Error and Mean Squared Error, with the ANN having a higher score on both. The purpose of OLS is to produce the model with the minimal MSE; however, this model is potentially overfitted due to the vast number of inputs, causing it to not generalise well to new data. However, in this scenario it seems to have been avoided as shown by the comparison graph of the unseen testing data.

AI, and in this case specifically regressor networks, are a useful tool in data analysis for building models of a system. ML can spot links in data that a human data analyst might not be able to see or may not be able to see as quickly.

Potential ethical issues with a project of this nature include a lack of desire to share usage information per electrical appliance in a household. This could be seen as an invasion of privacy as this data would tell you a lot about a person's lifestyle such as when they are usually in the house, when they sleep, etc. Furthermore, this data being publicly available would be a major security risk.

Future challenges faced in projects such as this include being able to predict the relevant data from less invasive inputs such as weather, time of day, world events, etc.

Bibliography

Galarnyk, M. (2024, February 23). *PCA Using Python: A Tutorial*. Retrieved from Built In: https://builtin.com/machine-learning/pca-in-pythonx

Rosidi, N. (2023, October 25). *A Beginner's Guide to Collinearity: What it is and How it affects our regression model* . Retrieved from stratascratch: https://www.stratascratch.com/blog/a-beginner-s-guide-to-collinearity-what-it-is-and-how-it-affects-our-regression-model/

TensorFlow. (2024, 04 24). *tf.keras.optimizers.SGD*. Retrieved from TensorFlow: https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/SGD