Energy Behavior and Patterns of Prosumers

Data Science & Analytics 5300: Statistical Learning

Kendall Gilbert, Hannah Kim, Powell Sheagren, and Irene Tait
GEORGETOWN UNIVERSITY

Introduction

As **solar panel installation by private entities** becomes more and more common, grid operators must grapple with the emergence of a new group of energy consumers: **prosumers**, households or businesses who both consume and produce electricity for the grid. While solar installations can cut down on business or household energy costs, **energy companies are often unable to flexibly respond** to hourly fluctuations in previously-stable sectors, leading to partial blackouts and or uncaptured energy.

Objective

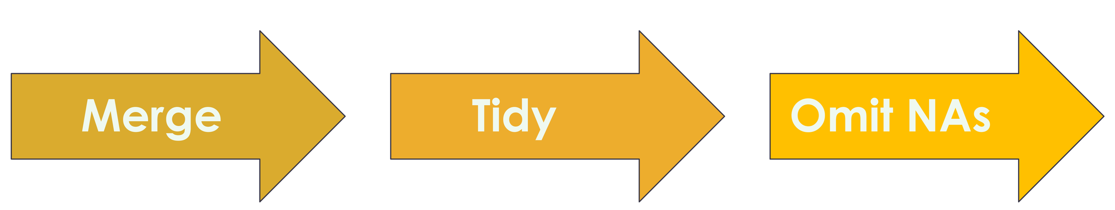
This project seeks to use statistical and machine learning tools to identify if there is a **trend in prosumer consumption** based on power production, current weather patterns, the price of both renewable and non-renewable energy, location, and time of day. Because businesses also **use energy differently** than households, we also attempted to classify if a given prosumer was a **business or household** based on their installed solar capacity, location, and net energy consumption vs production pattern.

Dataset

This dataset was gathered and compiled by Enefit Green on prosumers in **Estonia**. The dataset includes 6 CSV files, containing hourly weather data, demographic information on prosumer clients, **hourly energy prices** by type, and prosumer energy production and **energy consumption** in megawatt hours.

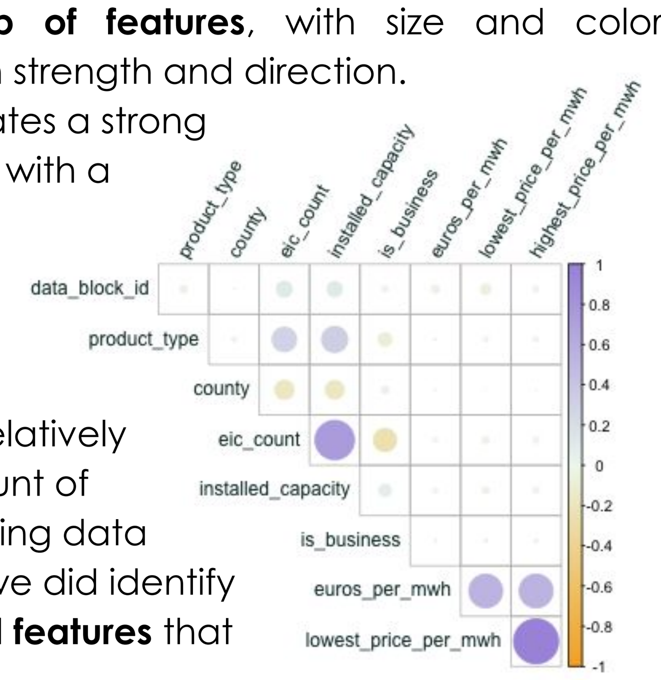
Enefit Green

Data Cleaning and Preparation



Exploratory Data Analysis

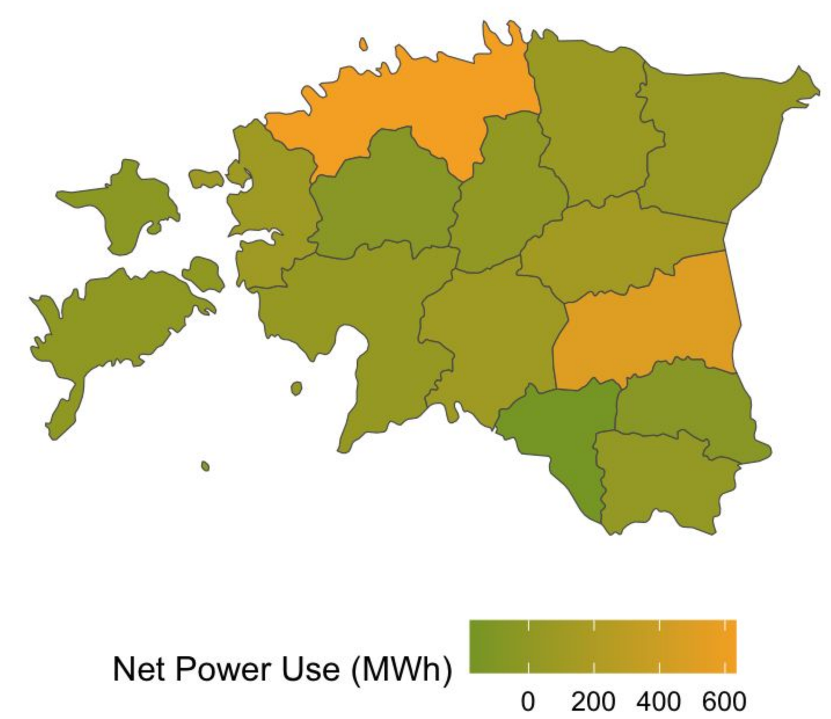
Correlation heatmap of features, with size and color indicating correlation strength and direction. A bright yellow indicates a strong negative correlation, with a bright lilac indicating a strong positive correlation.



Our Exploratory Data Analysis phase was relatively short, given the amount of familiarity gained during data cleaning. However, we did identify two **highly correlated features** that could be condensed.

Visualizing Energy Consumption

Net Power Usage Across Estonia
March 31 to May 30, 2023



Net power use calculated across county, subtracting total consumption from total production.

Statistical Methods

- Linear Regression with Ridge & Lasso
- Logistic Regression with Ridge & Lasso
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Neural Networks (Keras)

Regression Models

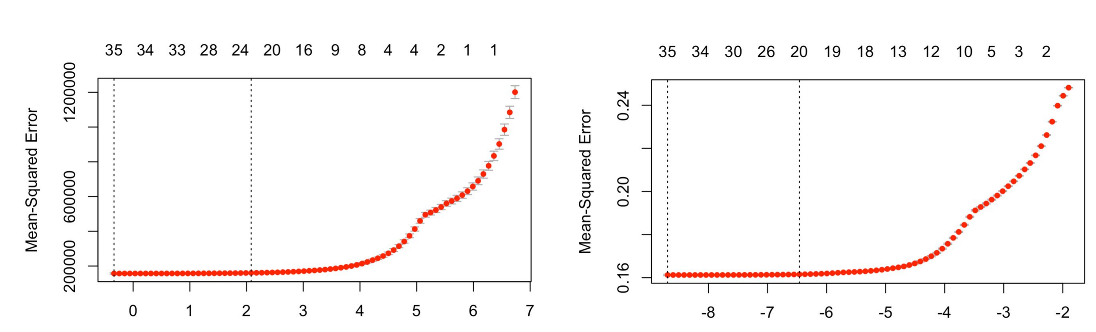
Logistic - Predicting whether prosumer **is a business or not**

Linear - Predicting amount of **net energy** that will be produced (in net MWhs) based on prosumer qualities

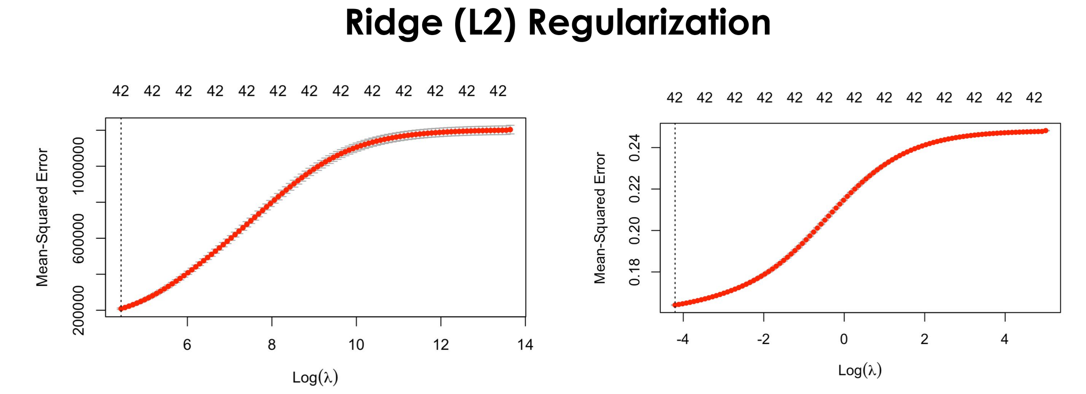
On both regression models, both Lasso (L1) and Ridge (L2) regularization were used to **reduce overfitting** and better **overcome multicollinearity**.

Linear | Logistic

Lasso (L1) Regularization

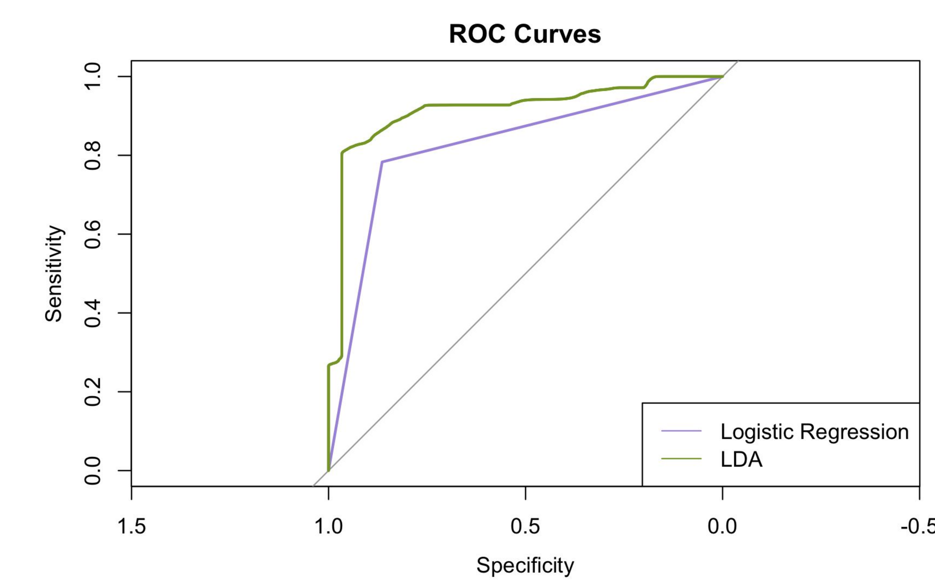


Ridge (L2) Regularization



Linear Discriminant Analysis

The LDA model performed with 86% accuracy, improving upon the logistic regression model's 82% accuracy rate.

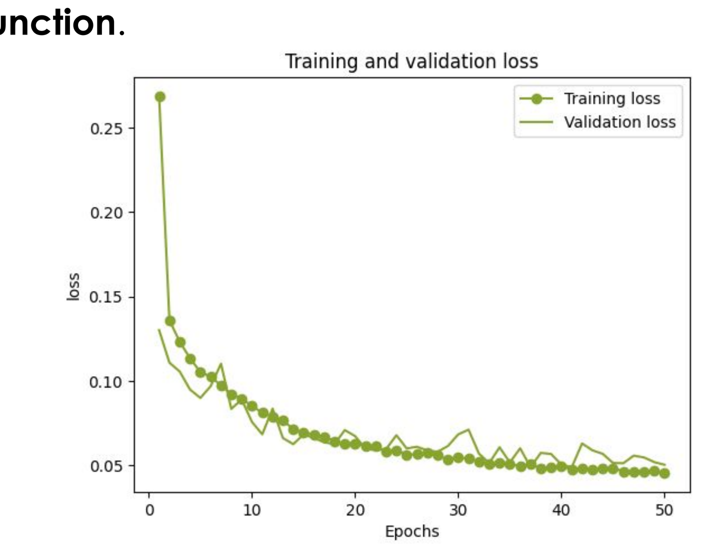


Artificial Neural Networks (ANN)

A **neural network** was applied to the data for both the regression and classification tasks after normalizing the data. 50 Epochs were computed with **early stopping**. Feature selection was not performed in order to reduce compute time, and **L1 and L2 regularization** were applied.

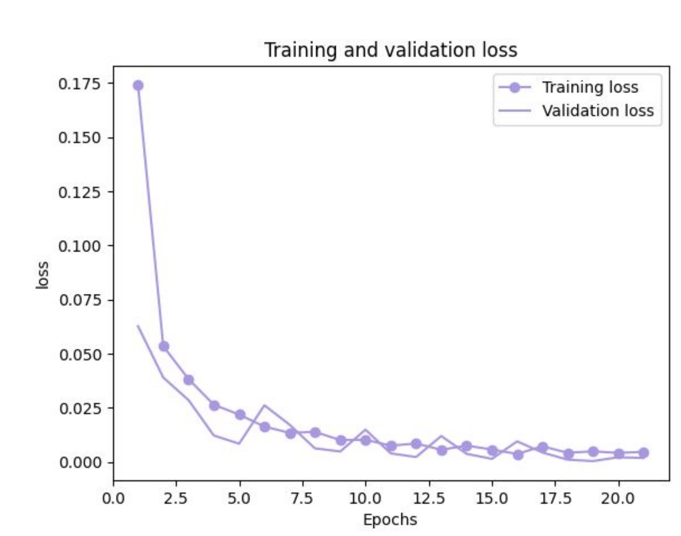
ANN Regression

The regression model used 1 input layer, 4 hidden layers, and 1 output layer. After hyper-parameter tuning, our final model contained the same layers, utilized **Adam** as an optimizer, **MSE** (Mean Squared Error) as a loss function, a dropout layer, **Relu** as an activation function, and a **linear output function**.



ANN Classification

The classification model used 1 input layer, 4 hidden layers, and 1 output layer. We again utilized **Adam** as an optimizer, a dropout layer, and **Relu** as an activation function, but used **binary cross entropy** as a loss function and a **sigmoidal** output function instead of linear.



Performance Comparison

Regression

Model name	Root Mean Squared Error
Linear Regression	395.75
Linear Regression + Lasso	396.06
Linear Regression + Ridge	456.44
Neural Network (Keras)	0.22

Classification

Model name	Accuracy
Logistic Regression	0.82
Linear Discriminant Analysis	0.86
Neural Network (Keras)	0.9995

Results

Regression Process

- Root Mean Squared Error**:** The **neural network** was the best method at approximating by large margins. This may be due to inherent non linearity in the data.

Logistic Process

- **Accuracy:** The **neural network** was the best method at approximating the value of whether or not a prosumer was a business. However, the **difference was not as great** as in the regression process, and since neural networks are more computationally expensive, a **discriminant analysis** model may be superior.

Conclusions

Our ANN model is able to **accurately predict the difference** between production and consumption of prosumers in a given area, **assisting grid operators** in understanding the behavior of this unique segment of their energy grid. We are also able to classify **whether a prosumer is a business or household**, which may help energy companies better segment prosumers based on energy behavior.

However, because all tested models rely on historical data, **changing weather and climate patterns** could lead to different energy needs and practices. Additionally, **policy decisions** that affect power usage or pricing could alter prosumer behavior. We are also intentionally interpreting several **time series variables as static** in order to fit the given models; while this does not invalidate these models, it limits the nuance that can be gleaned from their performance.

Future research could include **expanding into time series** modeling, exploring alternate nonlinear regression models, completing feature selection to highlight the most important variables, and **increasing the time of study**.

References

Kristjan Eljand, Martin Laid, Jean-Baptiste Scellier, Sohier Dane, Maggie Demkin, Addison Howard. (2023). Enefit - Predict Energy Behavior of Prosumers. Kaggle. Data accessed April 2023.
https://kaggle.com/competitions/predict-energy-behavior-of-prosumers