**Final Term Project - Predicting Household Electricity Cost Using Machine Learning**

Guruprasad Velikadu Krishnamoorthy

College of Science and Technology, Bellevue University

[DSC550-T302: Data Mining](https://cyberactive.bellevue.edu/webapps/blackboard/execute/courseMain?course_id=_512542_1)

Professor. Brett Werner

November 16, 2023

**Final Term Project - Predicting Household Electricity Cost Using Machine Learning**

1. **Introduction:**

**1.1. Narrative of the business problem:**

Budgeting and estimating the household expenses can be a big challenge for many. With soaring inflation, there are only a few avenues where one can cut costs, and utility expenses can be one of those if we understand our usage pattern and try to cut down on excess usage. But where to begin? In some states like Texas, hundreds of electricity providers offer a variety of plans that not only make it more confusing and very frustrating to consumers (Texas Electricity Companies | Best Electricity Supplier Texas, n.d.). To make matters even more complicated, some companies provide lucrative discounts for 12-month or 24-month contracts, which can become a huge trap if the consumers don't understand the billing pattern and are usually surprised by the hidden costs. Though the usage can vary from month to month, estimating the usage will give an idea of what to expect in the utility bill at the time of signing up. For instance, the residents in Texas may see a huge spike in their summer electricity bills mostly due to the extensive usage of Air conditioners to beat the summer heat. In contrast, the residents in Alaska may pay more for heating their homes. They can plan to allocate more in their monthly budget during excessive usage. Many other factors can contribute to electricity bills than just the AC or the heating equipment, which will be explored further in this project. ***Hence predicting the electricity costs for the household is the business problem that will be explored in this project.***

**1.2. Why is it important to predict the electricity costs?**

Apart from the budgeting reasons for the household, there are more compelling reasons for the energy companies to forecast and predict the spot and forward electricity prices. Since the de-regularization of the once monopolistic energy sector in many Western countries, the landscape of production and consumption of electricity has changed a lot. Adding to the complexity of electricity being a very special commodity, a continuous balance between production and consumption is required in the power grid as it is not very economical to store. A variety of factors can lead to price fluctuations such as precipitation, temperature, wind speed, etc. All these unique characteristics combined with seasonality can contribute to abrupt price spikes. Hence being able to predict electricity prices at a reasonable level of accuracy can help companies adjust their bidding strategies effectively thus maximizing the profits. (Electricity Price Forecasting, 2023)

**1.3. How would you pitch this problem to a group of stakeholders to gain buy-in to proceed?**

If there is one thing that customers hate the most regardless of any kind of business, it is the hidden costs. Unfortunately, many businesses don’t understand this and are instead more focused on churning profits, they soon come to realize that customer retention can be a very challenging deal. With many competitors entering the electric market, companies must think of innovative ways to retain customers. But where do they start and how can they achieve this? The solution is very simple, it is customer transparency. Being able to break down the utility bill into simple steps without too many complexities or additional (and hidden) charges can be key. This can help boost confidence and loyalty and can help with customer retention. Additionally, if the customers can be alerted about the potential bill, they can expect in the future months based on their usage pattern, it can take the customer confidence to the next level. Also, during times of extreme weather events such as excessive heat in summer or heavy snow in winter, there can be an extreme load on the power grid, and most of these costs are dumped onto the customers. If a machine learning prediction model can be deployed with reasonable accuracy to predict the monthly bill based on the usage pattern, it can make the company unique amongst the competitors, enhance customer stickiness, and lessen churning. ***Also in the future, it can be expanded to a Netflix-like subscription model with multiple tiers based on the usage pattern***. The subscription model has been proven very successful and has a great customer retention rate.  (What Will Electricity Pricing Look Like in 2040?, n.d.)

**1.4. Source of the data:**

The dataset for the project is taken from <https://www.eia.gov/consumption/residential/data/2020/index.php?view=microdata>. It is a collection of data taken from surveys collected from approximately 18,500 households in the US. These households statistically represent about one-third of the US population that occupy primary residences. The data set has about 18500 rows and 759 columns.

1. **Summary of Project Milestones:**
   1. **Exploratory Data Analysis:**
2. In the EDA phase of the project, we started by exploring the statistical summary of the dataset to identify the minimum, maximum, mean, and standard deviation of the numeric columns in the dataset. During this analysis, many categorical columns that were mislabeled as numbers were identified.
3. No nulls were found in the dataset, though there were many negative values in both the numeric and categorical columns.
4. Histograms and box plots were used to explore the distribution of the data in some numeric columns such as Total Square foot area, KWH, and home temperature, and found that most of them had outliers.
5. The bar plot was used to check the frequency of the distribution of items in categorical columns such as Wall Type, Equipment Type, etc.
6. A pair plot was plotted (figure 1) to study the relationship between various numeric variables and the findings are as below:
7. A positive correlation between KWH and DOLLAREL was found.
8. The Home temperature did not correlate with any of the fields in the plot.
9. The Groundwater temperature was found to be higher in southern states than in the Northeast or west.
10. The bigger the size of the household, the higher the electricity bill was, as a positive correlation was found between the two.
11. Most of the energy usage (KWH) in the dataset was from the southern states.A collage of data

    Description automatically generated

(Figure 1: Pair plot)

1. Finally, a correlation matrix was built, and the numeric fields that correlated with the target variable -DOLLAREL (electricity bill) were studied. A strong positive correlation between KWH and DOLLAREL was found. Also, columns such as Total Square feet, Number of Windows, Stories, Total rooms, etc., had a positive correlation with the target variable which was used in the further analysis of the project.

A screenshot of a chart

Description automatically generated

(Figure 2: Correlation Matrix)

**2.2. Data Preparation:**

The dataset had about 18500 rows and 789 columns. In the Data preparation phase, some of the low-hanging fruit such as column elimination and data transformations were first treated.

1. Unique columns such as DOEID, STATE\_FIPS, and columns with redundant information were dropped. Then the numeric columns with negative values were converted to zero and the negative quantities in categorical columns were converted to 99 to represent missing data.
2. A list of all the categorical features was identified and converted to dummy variables. Also, outliers identified in the EDA phase were treated by filtering them from the analysis.
3. After the initial elimination of the non-essential columns and after creating dummy variables, the dataset had a whopping 2655 features.
4. Then various data selection steps such as Thresholding features variance and mutual info regression were deployed, and the number of features was reduced to 23(shown in the figure below).
5. Finally, the Principal Component Analysis feature extraction method was applied, and the number of features was reduced from 2655 to just 13 features retaining 90% of the information.
6. The dataset was split into training and test sets before applying the PCA to avoid Data snooping issues during the model-building phase.

A screen shot of a computer

Description automatically generated

(Figure 3: Features after Feature selection)

**2.3. Model Building and Evaluation:**

As the target variable- DOLLAREL is a continuous quantity, a Supervised regression model was used to predict the value. Various methods were deployed to train the model as described below.

1. Multiple subsets of the dataset were used in each iteration of testing during which the dataset was split into train and test sets. The datasets were trained using 12 different kinds of Regression models in each iteration and the average of the results was compared to find the best model based on the R2 value. Based on this approach, Lasso, Ridge, and Ordinary Linear regression methods had the highest accuracy. The R2 value represents the goodness of the fit of the model and the metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were used to identify the best fit.

A screenshot of a black screen

Description automatically generated

(Figure 4: Results of Regression models )

1. In the hyperparameter tuning approach, various values of the hyperparameters of the regression models were tested using the GridSearchCV method using 5-fold cross-validation. The outcome of this approach was very similar to the first method and Lasso regression with an alpha value of 0.9 was found to have the best accuracy of 68.18%.

A screen shot of a computer program

Description automatically generated

(Figure 5: Results of Hyper Parameter Tuning )

1. Finally, an automated way of finding the best model using the PyCaret library was used results were very similar to the above two methods. The Simple Linear regression, Lasso, and Ridge models had the highest accuracy of about 68%.

A screenshot of a computer screen

Description automatically generated

(Figure 6: Results of Regression models using PyCaret )

**2.4. Summary of Changes made to the code after Milestone 3:**

**2.4.1. Milestone1:**

1. In the EDA (Milestone1), new refined plots for Histogram, Boxplot, and Bar charts with multiple subplots were added. The changes were made in the sections PLOT3, PLOT4, and PLOT5 of the Jupyter Notebook.
2. A new pair plot to represent the relation between variables was added in section PLOT6.

**2.4.2 No changes were made to Milestone 2**

**2.4.3. Milestone 3:**

1. A new polynomial feature of ‘degree 2’ was added during the model-building phase which improved the model accuracy slightly.
2. Also in the same step, a new logic of running multiple iterations of different dataset samples to train the 12 types of regression model was added. The average of the results was then calculated from each iteration. Please note this is different from Cross validation which was also deployed in the Hyper Parmeter Tuning step.

A screenshot of a computer program

Description automatically generated

(Figure 7: Code snippet highlighting the changes made )

**3. Conclusion:**

**3.1. What does the analysis/model building represent?**

The various regression models trained in this project had an accuracy rate of about 68% which indicates about 68% of the variability in the Target feature - Electricity cost can be predicted by the features in the model. RMSE (Root Mean Squared Error) and Mean Absolute Error (MAE) were used as the metrics to evaluate the model which represents how close is the predicted value to the actual value. Lasso and Ridge models have an MAE of ~300 and an RMSE of ~450. The RMSE is greater than MAE which may be due to the presence of some large errors in the dataset. This indicates that on average we can see a difference of about $300 to $450 in the electric bill that is predicted by the model.

**Is the model ready to be deployed?**

Before deploying a new model in production, it is essential to know how the new model is better than the existing one. Though it is not very simple to conclude if our model is doing better, comparing the performance with the already existing model can help us determine if the model is ready to be deployed.

If there are no other existing models, a ***shadow deployment*** may be a good choice. In a shadow deployment, the performance of the entire pipeline with the production-like data can be tested, and if we find any issues the model can be revisited to make enhancements. As the accuracy of the model is about 68%, the model can be deployed as shadow deployment and assess how the model performs before deploying it in the live production environment and see if the accuracy improves.

**3.3 Recommendations:**

The dataset used in this model has about 18,500 rows that statistically represent one-third of the US population that lived in Primary residences in 2020. As energy prices are largely impacted by Inflation, testing the model with a more recent dataset with recent energy prices is recommended. Also, the more the data, the model can train with more samples. Hence collecting more data samples is another recommendation. These factors may also help with improving the accuracy of the model.

**3.4 Potential Challenges or Additional opportunities to be explored?**

The dataset did not include features such as temperature, air pressure, humidity, wind speed, or any other weather-related metrics that can have a huge impact on predicting electricity use. Also collecting the data over a period and building the model with period as a feature can be another opportunity to be explored. The current model mostly relies on the KWH data to predict electricity costs. But as an additional opportunity, the model can be expanded to predict the usage and the final bill with the inclusion of demographic features such as the population of the city, terrain conditions, etc. that may help improve the efficiency of the model.

**References:**

*Texas Electricity Companies | Best Electricity Supplier Texas*. (n.d.). Www.texaselectricityratings.com. Retrieved November 14, 2023, from https://www.texaselectricityratings.com/companies

‌*What will electricity pricing look like in 2040?* (n.d.). Utility Dive. Retrieved November 14, 2023, from https://www.utilitydive.com/news/what-will-electricity-pricing-look-like-in-2040/558708/

*U.S. Energy Information Administration - EIA - Independent Statistics and Analysis*. (n.d.). Www.eia.gov. <https://www.eia.gov/consumption/residential/data/2020/index.php?view=microdata>

*Electricity price forecasting*. (2023, August 13). Wikipedia. <https://en.wikipedia.org/wiki/Electricity_price_forecasting#:~:text=A%20power%20market%20company%20able>

‌