August 7, 2024

C964: Computer Science Capstone

Task 2 parts A, B, C and D

Part A: Letter of Transmittal

August 7, 2024

Energy Commissioner
Oakland County Energy Commission
123 Main St.
Novi, MI 48375

Dear Energy Commissioner,

With the rising energy usage across the globe, having an effective energy consumption forecasting solution is crucial to manage resources efficiently and maintain energy grid stability. I propose a machine learning solution to enhance our energy consumption forecasting capabilities within Oakland County. This solution will optimize energy distribution, increase efficiency, and support the continual need for sustainable energy practices.

Accurately forecasting energy consumption has become more challenging as technology advances. Existing methods rely too much on historical data and simple models, often not including dynamically changing events and energy consumption patterns. These existing methods result in inefficiencies, higher costs and risks, and usually energy shortages or surpluses that strain the community's infrastructure.

To address this problem, our organization proposes implementing a machine learning-based forecasting system. With our state-of-the-art machine learning algorithms, we can ingest and analyze a tremendous amount of data to identify specific patterns that lead to higher accuracy and reliability of forecasts. Through integrations with various data sources, including historical consumption data, weather patterns, and economic and event-driven trends, our solution can provide real-time actionable forecasts into the county's energy needs.

Implementing the machine learning solution will benefit the county by:

1. Enhancing the precision of energy consumption forecasting for better planning and resource allocation.
2. Enable more efficient energy distribution with improved precision, reducing the likelihood of shortages and increasing stability.
3. Potential for lower energy prices for county consumers with the improved energy efficiency.

The solution will follow a timeline consisting of:

Phase 1: Data Collection and Preparation – We gather the relevant data from approved sources and clean it to be processed.

Phase 2: Model Creation and Training – We develop and train the machine learning model using the data gathered and cleaned.

Phase 3: Testing and Validation – We test the model from the previous phase to ensure accuracy and reliability.

Phase 4: Deployment and Support – We provide training to the county personnel and deploy the solution. We will use monitoring tools for continuous improvement.

The estimated cost for this project is $227,280, which covers the development tools, cloud hosting, development team, and data acquisition. Oakland County, being a public entity, has additional funding options such as grants, and other public/private partnerships can be explored to mitigate financial impacts to the county.

Our team consists of leading machine learning, data science, and energy system professionals with a proven track record of implementing successful similar projects for small and large municipalities. Our solutions have both proven to decrease the financial strain of energy consumption and improve sustainability leading to a greener world.

This project is a significant opportunity for Oakland County to lead surrounding municipalities in energy innovation and sustainability. I am confident in our organization's ability to improve energy forecasting and management capabilities.

Sincerely,

Robert Kearns

Robert Kearns
Director of Technology – Machine Learning
K-Tech Industries

Part B: Project Proposal Plan

Project Summary

Oakland County currently faces challenges in accurately forecasting energy consumption due to the increasing complexity and variability of the technology we use in the 21st century. The county faces inefficiencies, higher operational costs, and potential energy shortages with the current methods. The primary client is the county's energy commission, which will use the forecasting models to enhance the planning and resource allocation for better energy consumption. The commission aims to improve operational efficiency, reduce costs, and support sustainable energy consumption practices using machine learning tools and models.

The deliverables will include:

- Energy Consumption Forecasting Model: The machine learning model will be able to predict and forecast energy consumption with a high degree of accuracy.
- User Interface: An intuitive dashboard will be available for the stakeholders to review the historical data and visualizations and perform predictions of energy consumption based on the data.
- Documentation and Analysis: A comprehensive user guide detailing the use of the application, which will include step-by-step guides for the model setup and configuration.

With the machine learning application, the client can provide a more accurate forecast of the county's energy consumption, leading to better planning and resource allocation. The application will also have economic advantages for all county members by lowering the operational costs, improving energy efficiency, and reducing energy prices for consumers.

Data Summary

The raw data for the machine learning solution will be collected from the county's energy commission, specifically the historical data. Along with this, data will be collected from energy management companies in the area, such as PJM Interconnection LLC, for additional training and simulation for the model.

For the initial training of the machine learning model, the data will be collected in individual CSV files to be preprocessed and cleaned. This cleaned data will be given to the development team to create the machine learning solution tailored to the county's needs. If anomalies or missing data exist in the dataset, techniques such as normalization, imputation, or data transformation will need to occur before the final solution is deployed. Once the data structure and model have been finalized, a real-time and automated system will collect data continuously directly from energy sources.

The selected data that will be collected from the county's energy commission of PJM Interconnection LLC are comprehensive and directly relevant to the county's needs for enhanced energy consumption forecasting. There are no obvious ethical or legal concerns regarding the data. The data obtained and passed through the model will be anonymized as much as possible in compliance with relevant privacy and data protection guidelines.

Implementation

The implementation and development of the machine learning solution will follow the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. The phases of implementation will be as follows:

- Business Understanding: An investigation will be conducted to understand Oakland County's needs to optimize energy distribution and reduce operational costs through machine learning-enabled forecasting. At this phase, we will determine the business objectives, assess the current state of energy systems and historical data trends, define forecasting accuracy goals and metrics, and define the project plan for timelines and required resources.
- Data Understanding: At this phase, the data sources will be identified, collected, and an initial analysis will be performed to understand the data fully. This phase will include the initial data collection of historical energy consumption data. This data will then be analyzed to understand the structure and properties applied for it to be ingested into the machine learning model. Along with the analysis, an assessment of the data quality will be performed to identify missing values, inconsistencies, and anomalies to be addressed.
- Data Preparation: In this phase, the data will undergo a cleaning process to handle any missing values or inconsistencies that are found. For the datasets obtained through different source systems, the data will be combined into a single monolithic dataset and then transformed into a format that is ingestible into the machine learning model for training. At this phase, relevant data features will be selected based on their importance to the county and bolster forecasting accuracy.
- Modeling: We will use the collected and clean data to develop and train the machine learning solution, specifically the Random Forest Regressor algorithm. During the training, we will include test sets for real-time testing of the model.
- Evaluation: During this phase, with assistance from the stakeholders, we will evaluate the model's accuracy and performance. This is to ensure it aligns with business objectives and industry standards. During the evaluation of the output, we will measure the MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and RMSE (Root Mean Squared Error) scores. Using these scores, we will present our findings to the core Oakland County implementation team to assess the progress and determine if the solution is ready for deployment.
- Deployment: We will deploy the solution using a cloud-based technology to compute and store the data, such as Microsoft Azure. Deployment will also include a web-based interface for users to view the compiled historical and forecast data. During deployment, user documentation and training will also be provided to the Oakland County Energy Commission team members on how to use the web interface and interpret the results of the solution's forecasts.

Timeline

Milestone or deliverable	Duration	Projected start date	Anticipated end date
Initial setup and planning of the project team. Obtain project approval by project sponsor. Begin prototyping of solution.	15 days	09/01/2024	09/15/2024
Identify all relevant data sources. Clean and process the data to be ingested into the machine learning model. Complete first round of prototyping and create the AI framework of Random Forest Regressor.	45 days	09/16/2024	10/31/2024
Create the automated pipelines for data ingestion and model training. Begin and evaluate the model's predictions	28 days	11/01/2024	11/29/2024
Complete the model's training and confirm evaluation of the model meets stakeholder and industry standards. Begin dashboard creation. Test cloud connectivity and access.	51 days	12/02/2024	01/22/2025
Deployment of solution and obtaining final sign-off from project sponsor. Create employee accounts within the dashboard and train personnel in the usage of the model and how to interpret output.	13 days	01/23/2025	02/05/2025

Evaluation Plan

During each stage of development, there will be assessments and validations to ensure the accuracy of the data product. There will be an initial data quality assessment performed upon the collection of data where the team will identify missing values, outliers and anomalies. Throughout the development process, ongoing data monitoring will be performed through automation scripts that will monitor incoming data for quality. If there is an outlier in the standard that is defined, an alert will be triggered to the development team for further investigation. A standard cleaning process will be implemented such as data imputation for missing values and data normalization to maintain data quality throughout the project.

The machine learning model will include verifications at each stage of development as well. This will include performance metrics based on Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to assess the model's accuracy and reliability. This will be applied at regular checkpoints throughout the model's development and training to make adjustments as needed throughout the project.

The code used to development the machine learning solution and infrastructure will also go through verifications. This will include code reviews held by the development team to uphold industry best practices and identify potential issues early. Unit testing will also be performed for each component of the solution to ensure they function as intended as individual packages. Additional verifications will be

performed on the integrations with source systems for automated data collection and infrastructure stability.

Upon completion of the project, numerous validation methods will be performed. The machine learning model will have a separate validation dataset that was not used during testing that will serve as an unbiased test of the model's forecasting abilities. The model will also be validated against the existing forecasting methods of the county to showcase improvements made and justify the solution's capabilities.

Before the solution is deployed into production, User Acceptance Testing (UAT) sessions will be conducted with select members of Oakland County's Energy Commission. The members will include key energy management officials and IT staff to ensure the application meets their needs and expectations. During this time, the project team will collect feedback to make any adjustments prior to go-live.

After the solution is deployed into production, the project team will assist the county with continuously monitoring the performance of the solution for accuracy and reliability. The project team will use the forecast data that is output from the solution and compare it with the actual data collected by the county. Surveys will be given to members of the Oakland County Energy Commission team to ascertain satisfaction of the solution and implementation process.

The quantitative metrics to be measured include:

- Accuracy: A target accuracy rate of at least 90% correct in forecasting energy consumption.
- Error Rate: Calculated MAE will be less than 0.1 and RMSE will be less than 0.15 to maintain prediction accuracy.
- Throughput: Measure the number of predictions that the solution can make in a defined timeframe (second/day/month).

Resources and Costs

Resource	Description	Cost
Development Tools	IDEs (Visual Studio Code) and machine learning libraries (Scikit-learn, Pandas, Matplotlib and Seaborn, Plotly)	$0 (Open-source)
Cloud Hosting	Azure Compute Instances (Virtual Machines) and Azure Blob storage	$2,500/month $500/month
Data Sets	Oakland County historical energy data, PJM Interconnection LLC, CBECS and LL84 Data,	$0 (publicly available)
Development Team	Data Scientist, 2 Machine Learning Engineers, Project Manager, 2 Software Developers	$7,000/month, $14,080/month, $4,000/month, $6,800/month
Miscellaneous Costs	Office space, travel, training, and incidental costs	$3,000/month
	Total	$37,880/month 6 month project: $227,280

Part C: Application

Required files:

- **main.py**: Contains the source code for the machine learning solution
- **PJME_hourly.csv**: The simulation dataset that is used to train and predict the energy consumption
- **C964-Capstone-Kearns1.ipynb**: The Jupyer Notebook file made available and loaded into the Google Colab link below.

Links for evaluation:

- https://colab.research.google.com/drive/196c44bvLMJwhJtv0EOe5xgs9g8Bd9WTW?usp=sharing

Part D: Post-implementation Report

Solution Summary

Oakland County and its energy commission faced challenges in accurately forecasting the energy consumption rates that increase with complexity as we move toward a more technological world. These challenges rise due to the complex and dynamic nature of the energy usage patterns within the municipality. The solution to this problem was implementing a machine learning-based application using a Random Forest Regressor algorithm model to enhance forecasting accuracy and efficiency.

The deployed solution collects energy consumption data from the county and other sources, analyzes the patterns and trends, and then utilizes machine learning technologies to provide precise and actionable forecasts for improved energy distribution, reduced energy-related costs, and moves the county towards its sustainability goals.

Data Summary

The raw data was collected directly from Oakland County's Energy Commission to ensure we had the most relevant data to train our models. We also collected data from energy management companies in the area, such as PJM Interconnection LLC, for additional training and simulation for the machine learning model.

Each data source had to be cleaned and processed to ensure integrity and handle missing values, outliers, and inconsistencies. Some techniques that were used during the cleaning process were imputation, normalization, and data transformation. As the data was obtained before deployment and the ongoing collection of data, it is stored in a secure cloud-based database that can scale as time goes on and performs regular backups for disaster recovery best practices.

Machine Learning

The machine learning method chosen for the solution was the Random Forest Regressor algorithm. This type of method is typically used for regression-based tasks where the concept involves constructing multiple decision trees, creating a forest during the training of the model, and then outputting the average prediction of the individual trees.

The method was developed using the Python programming language and the scikit-learn library. We began the process by loading in energy consumption data to help us simulate the prediction model and perform initial trend analysis. With the raw data, we extracted additional features from the Datetime index, such as year, month, day, and hour. For training and preparing the test datasets, we split the data into an 80-20 factor; 80% of the data was used for training the model, while 20% was saved for testing. We could predict energy consumption with the trained model, and the predictions were compared to actual values for validation.

Many algorithms and methods were viable for the problem of accurately forecasting energy consumption for a given area. The Random Forest Regressor method was chosen for its ability to handle large datasets and its performance capability in regression tasks. The method also proved to have a stable and accurate

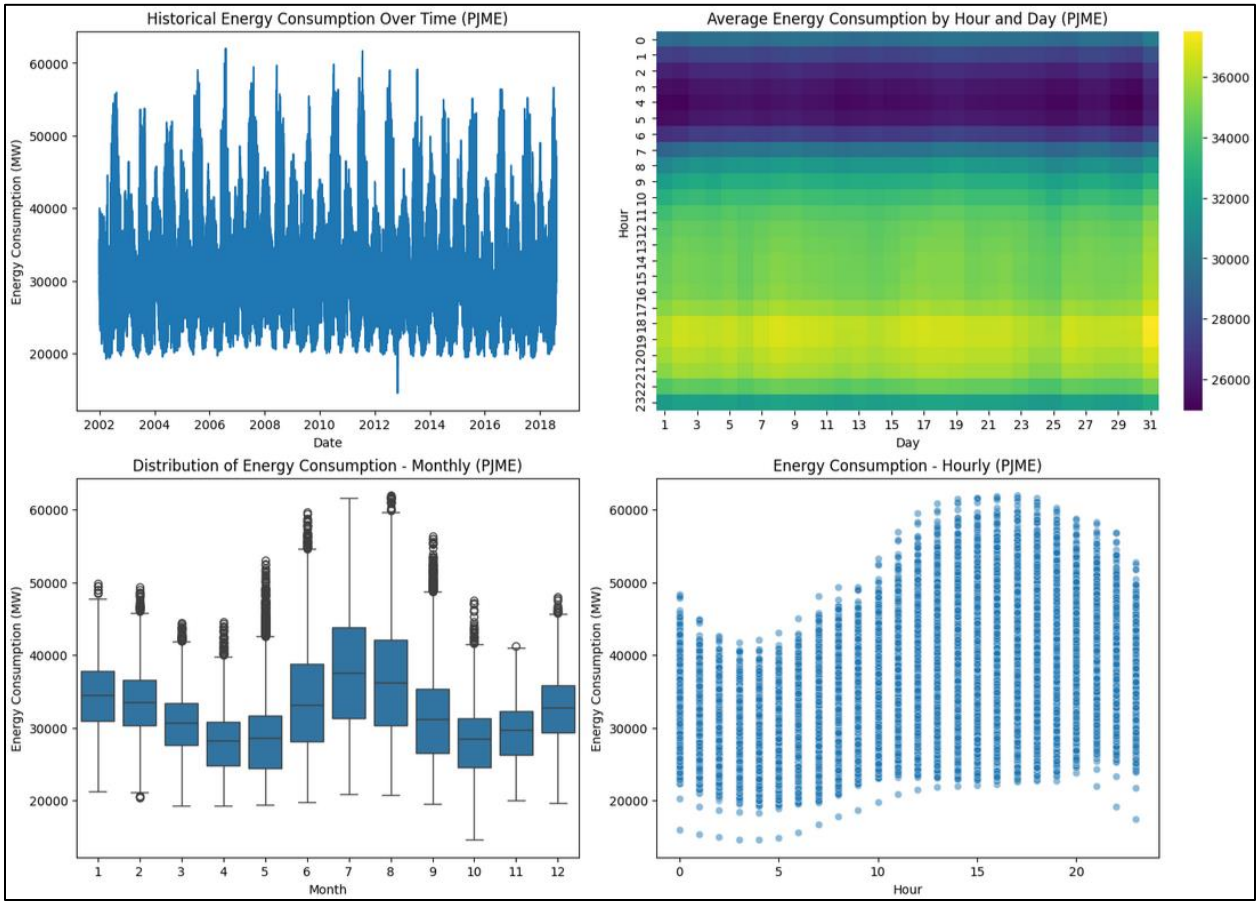
prediction, especially that energy consumption data can potentially have a wide range of values and influences such as weather, special events, and natural disasters.

Validation

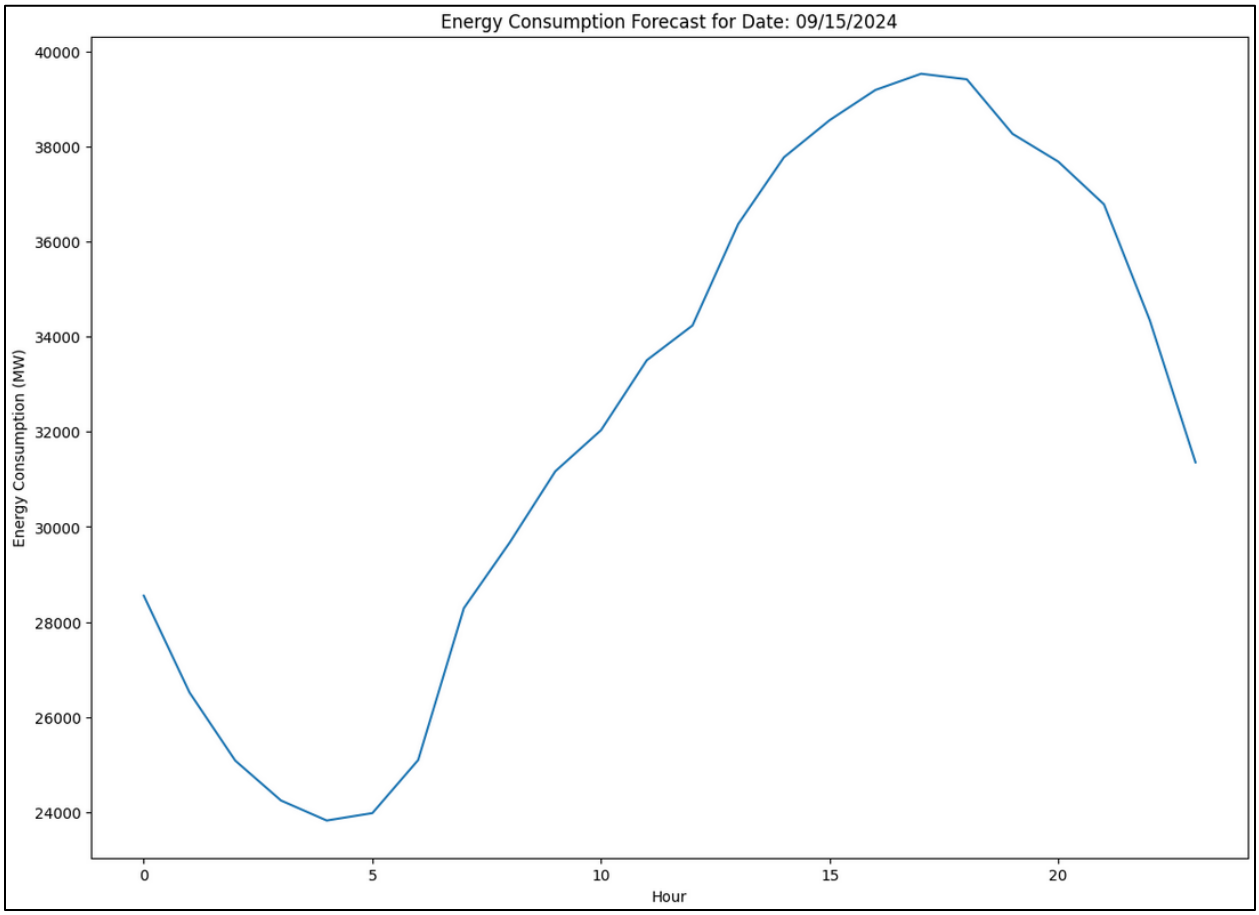
A Hold-Out Validation method was used to ensure the model and predictions held up to industry standards. In our solution, we separated the simulation data into two sets by a factor of 80-20, where 80% of the data was used for training the model while 20% was saved for testing. This splitting of data allowed us to use actual data to compare and measure the error rates of our predictions. The results of these validations concluded that the Mean Absolute Error (MAE) was 1,050.15 MW with a percentage error of 1.69% - 7.22% based on the minimum and maximum values in the simulation dataset. The Root Mean Squared Error (RMSE) was 1,655.21 MW with a percentage error of 2.67% - 11.38% based on the minimum and maximum values in the simulation dataset. These results of the MAE and RMSE confirmed that the model provided accurate and reliable results in predicting energy consumption that is well within the industry average.

Visualizations

The application provided multiple unique visualizations for observing trends in historical energy consumption data and viewing energy consumptions forecasts. The types of visualizations that were used to demonstrate trends were a line plot, heat map, box plot, and a scatter plot.



In addition, a lone plot was used to visualize the energy consumption forecast for the user provided date



User Guide

The intention of the machine learning solution was for it to be easy to use even if the user is not technical or understands code. Before running each block of code, read the informational text above it to understand what the code is doing and if there are expected inputs or outputs. When the instructions instruct to run a code block, the user can either click within the code block and press shift+enter on their keyboard or hover their cursor over the code block and select the play icon in the top-left corner of the code block section.

1. Navigate to the following Google Colab link: https://colab.research.google.com/drive/196c44bvLMJwhJtv0EOe5xgs9g8Bd9WTW?usp=sharing
2. Select the folder icon and then the upload file icon on the left side of the page,
3. Upload the provided dataset named “PJME_hourly.csv” to the application.
4. Run the first block of code. This will import the necessary Python libraries into the session.
5. Run the second block of code that will load previously uploaded dataset to the session and read the CSV data. The output will contain a preview of what the raw data looks like in a tabular

format. It will then transform the data to obtain additional features for the machine learning model. The output will contain a preview of the transformed data.

6. Run the third block of code. This will create the Random Forest Regressor model and then build the decision trees based on the test data from the uploaded dataset.
7. Run the fourth block of code. This will perform multiple calculations such as finding the minimum and maximum values in the dataset, calculating the Mean Absolute Error and percentages, and calculating the Root Mean Squared Error and percentage. The block of code will output the computed values.
8. Run the fifth block of code. This takes the data found in the uploaded data set to produce four visualizations: a line plot, heat map, box plot, and a scatter plot based on different time frames. The output will be a single image divided into four quadrants containing the graphs.
9. For the sixth block of code, the user will input a date in the future in the format MM/DD/YYYY within the quotation marks. For example, if the forecast is required for September 15th, 2024, the user would put 09/15/2024 into the quotation marks. Once the date has been typed, run the block of code.
10. Run the seventh block of code. This will take the date that the user entered previously to predict the energy consumption of that day based on the trained machine learning model. The output will be a table containing the megawatt usage per hour that day.
11. Run the eighth block of code. This will produce a visualization in the form of a line plot showing the user the forecasted energy consumption of the day entered.

Reference Page

No outside references or sources were used for this document.