Geothermal, oil and gas well subsurface temperature prediction employing machine learning.

Ameya Kshirsagar * and Parth Sanghavi
Department of Computer Science and Information Technology, Symbiosis Institute of Technology, Pune, Maharashtra, India
*ameyark28@gmail.com

Keywords: Support Vector Regression; XGBoost; SMU; GDR; AASG; Wells; Geothermal; Oil; Gas.

ABSTRACT
Geothermal energy is getting more and more attention these days due to its nature of being a clean source of renewable energy sources provider with a zero-carbon footprint, free and available year-round. Assessment of this geothermal system is often neglected in practice during the exploration phase of energy extraction in geothermal development projects. The artificial intelligence expertise encompasses a vast number of deep learning and machine learning techniques ranging from linear regression, logistic regression, convolutional neural networks, genetic algorithms, reinforcement learning experiments, generative adversarial networks, etc. On amalgamating these available machine learning techniques with the geothermal concepts like power conversion, geology, thermodynamics, geophysics, electricity generation, heat flux, etc. are capable of improving the geothermal sector holistically by Subterranean resource characterization, Discrete micro-seismic event detection, and classification, drill fault deception, and prediction, predicting subsurface temperature and geothermal gradient and more. This study targets the analysis and prediction of the subsurface temperature of geothermal oil and gas wells. The implemented Support Vector Regression (RMSE = 4.75) shows better results when compared with the counter model, XGBoost (RMSE = 5.16).

1. INTRODUCTION
Bottom-hole temperature (BHT) readings were increasingly utilized in the United States to map subsurface temperatures levels for geothermal resource analysis. In 2021, (Shahdi et al., 2021) combined BHT data from the northeastern U.S. with stratigraphic information & generated surface heat flow and temperature-at-depth maps using a basic thermal conductivity model. (Teresa Jordan et al., 2016) undertook a comprehensive review of the risks and potentials connected with probable geothermal resources in Pennsylvania, New York, and West Virginia. Each geothermal system is unique in its basic features, including the reservoir rocks, clay cap, fluid, and heat source. These variations are directly tied to the methods by which each type of geothermal energy is generated and the eruptions & modification of hydrothermal activities. These factors (Sutarmin & Daud, 2021) significantly affect the features of the rocks created, including their permeability, saturation, density, porosity, and the clay minerals contained inside them.

The energy stored in low-temperature geothermal regions in the northeast can be exploited for a range of direct-use applications. However, most geothermally active regions are located in the western United States (along Earth's plate borders). While it has been demonstrated that geothermal energy may be used for a variety of industrial and home direct-use purposes, few geothermal sites in the northeastern states are financially vulnerable. Two fundamental geothermal properties, heat flow and temperature at depth have been extensively explored using physics-based models. The temperatures of the reservoirs have a significant effect on a geothermal system Drilling will be extremely accurate for estimating the temperature of the geothermal reservoir, in contrast to the geothermometer and the geochemical technique. The presence of a rock resistance-reservoir temperature connection, which is commonly connected with TOR (Top of Reservoir) data gathered from boreholes and BOC (Best of Conductor) data generated from 3D M.T. data, is unquestionably significant to uncover.

Additionally to the geothermal energy sector, the subsurface temperature is an enormously important metric in the petroleum sector and other industries. Since hydrocarbon characteristics are highly influenced by temperature, they must be estimated before being utilized in reservoir & drilling models. Geothermal gradient maps are commonly used to acquire the geothermal gradient value at the required location and then compute the subsurface temperature at the desired depth of interest using the geothermal gradient value (Jones, 1970). Investors may make more confident judgments with the help of machine learning & geostatistics, which have been applied in a range of applications. Given the inaccessibility of geothermal energy, there is also a significant level of risk and uncertainty connected with the search drilling and geothermal energy production. Only a few extensive studies have been conducted to examine the hazards associated with geothermal development to give insights into the possibilities of developing geothermal locations. Machine learning is a new technology that has benefited the sector during the many stages described for geothermal energy.

2. LITERATURE REVIEW
In renewable energy, optimizing geothermal plant power production & economic value over years of operation is a key problem. To get the best results, one needs to predict the flow rates & output temperatures of production wells depending on the injection wells' input temperatures and the system's previous data. Based on previous data, (Carneiro et al., 2021) investigate the potential of machine learning systems to predict temperature outputs accurately. They propose an alternative approach, considering the difficulties associated with obtaining an evidence-based set of data from field measurements large enough to enable reliable ML. Attempting to develop a high-fidelity reservoir model & utilizing computational resources to generate a dataset large enough to enable reliable ML. It is presented that
Kshirsagar and Sanghavi

preliminary findings from applying machine learning to forecast the temperature time series of basic modeled geothermal systems will help them achieve this aim in the first stage. To extract temporal patterns from model data, they discuss the use of appropriate state-of-the-art machine learning algorithms, like Long Short-Term Memory networks or Convolutional Neural Networks.

Geothermal scientists were able to construct heat flow and temperature-at-depth maps using bottom-hole temperature measurements from petroleum well records, which have been used to detect possibly geothermally active areas. The use of different machine learning algorithms for predicting temperature-at-depth or geothermal gradients parameters is examined in this paper since there are many uncertainties and simplifying assumptions associated with the current state of physics-based models. (Shahdi et al., 2021). Their exploratory investigation discovered that XGBoost & Random Forest are the two models that forecast subsurface temperature with the best accuracy possible. They also apply their model to areas surrounding the sites to generate 2D constant temperature mapping at three different depths using the XGBoost model, identifying potentially geothermally active locations. As a benefit, they used an additional dataset including recorded temperature reading throughout the depth of 58 wells located in the state of West Virginia to evaluate both the suggested XGBoost & DNN models. In terms of accuracy, ML methods are generally equivalent to the physics-based model & in some instances, surpass the thermal performance model because the absolute mean error of Machine learning algorithms like XGBoost & DNN is 7.3, 7.27, whereas the physics-based model error is 8.76.

The conductivity of rocks and the reservoir temperature are mostly affected by the same properties, such as porosity, permeability, fluid salinity, and temperature. The goal of this study (Namaswa et al., 2021) was to combine temperature readings from Olkaria Domes geothermal well drilling with geophysical electromagnetic resistivity data to develop a new method for predicting reservoir temperatures via machine learning analytics. The Data-Driven Discovery Predictive Model Algorithm was created in Python utilizing the Anaconda framework. For coding and visualization, the open-source online tool Jupyter Notebook was utilized. The regression techniques used were: Decision Tree Regression, Adaptive Booster Regression, Random Forest Regression and Support Vector Regression. The R-Score & Mean Absolute Error metrics evaluated the model’s performance. Based on these performance scores, the optimal model for predicting subsurface temperatures from resistivity was recommended. The DTR algorithm improves outcomes with an R2 of 0.81 and an MAE of 29.8. (Chaayer et al., 2021; Kshirsagar, 2018) In high-temperature hydrothermal environments, the DTR approach can measure the subsurface temperatures from resistivity.

(Rundle et al., 2016) The authors are particularly interested in determining whether any predictors of micro-earthquakes can be used to forecast when significant seismic activity will occur soon, allowing for immediate action being taken before anything violent occurs. With an emphasis on characterization rather than detection, an unsupervised learning approach can be used to deploy machine learning techniques that forecast geothermal heat flux results based on variables of geologists. It includes bedrock, topography, crustal thickness, and other geologic variables for a given area — as well as geothermal heat flux information from around the world. Associating template fingerprints with distinct fault attributes or processes will lead to supervised learning approaches. These new technologies will help reservoir engineers to make better decisions, minimize harmful seismicity, and optimize small-scale seismicity to improve fracture networks and energy production efficiency. Identifying these processes and their transitions will substantially increase our knowledge of a geothermal reservoir's thermal-mechanical condition. These approaches also give a novel tool to discover and quantify spectrum variations in micro-seismicity, which may help us better understand the earthquake process.

Temperature and pressure data obtained throughout the wellbore are used in geothermal wells log analysis to estimate feed zones, reservoir pressure & reservoir temperature. The use of a multi-layered convolutional neural network to diagnose sets of temperature and pressure well records is discussed in this research (Okoroafor et al., 2021). This approach allows for the interpretation of many wells logs in seconds. This project's data source is synthesized well data that strongly match real data. A total of 10,000 datasets were utilized in this study. The data were partitioned into three sets: training data, validation data, & testing data: 8,000, 1,000, & 1,000. The algorithm takes three “depth-series” logs of temp, pressure, as well as temperature gradient as input, runs the data through it with a convolutional neural network with a flatten layer and a fully connected layer, and outputs five variables: reservoir temperature, feed zone depths, reservoir pressure, and depth at which reservoir pressure is known. This model's cost function was mean squared error. Adam was the optimizer method utilized, and the learning rate decayed exponentially. The algorithm recorded the model state with the least overall absolute validation error. With a TensorFlow backend, the architecture was built in Keras. The best model discovered throughout the hyper-parameter tuning procedure was utilized to forecast reservoir properties for the testing and validation data sets. With a training error of 0.8 percent, a validation error of 2%, and a test error of 3%, the findings demonstrate a good match between projected and actual data.

3. RESEARCH METHODOLOGY

EDA is a process of generating statistical information for numerical data in a dataset and constructing different graphs and charts to understand the data better. Geothermal energy may be a valuable source of heat energy for residential areas at shallow depths. Geothermal wells were drilled for production when the potential geothermally active zones had been identified. The drilling stage of a geothermal project might cost up to 45% of the entire project cost. ML has benefited the business in designing this stage more effectively from several perspectives. In the oil and gas sector, the EDA approach will be utilized to forecast geothermal subsurface temperature visually. This article uses mathematical models to conclude geothermal temperature forecasts based on data that has been visually reviewed.

The data selected is a combination of wells observation from SMU's (National Geothermal Data System) node & respective state observations collected from the (Association of American State Geologists) and GDR. The database provides information on wells identity (Observation URI, Well Name, and Well Status); county, state, latitude/longitude, the spatial reference system (SRS) that used map points, the depth of temperature measurement, temperature, the information source, and the parent dataset (e.g., SMU or AASG). The entire dataset was filtered for fewer temperature criteria, with temperatures ranging from 30 °C - 150 °C. Total 306,608 rows were there in dataset, out of which 214,997 rows are not null. It is better to look at and compare the results of each data collection using a variety of
exploratory methodologies. This procedure uses visual & quantitative tools to analyze the dataset and locate missing data and outliers. It suggests reasonable next steps, queries, or research topics for future research.

3.1 EDA Workflow
EDA involves summarizing data numerically & visually and, as a result, preparing data for even more formal modeling procedures. EDA can quickly deliver useful information, identify patterns, and uncover overall relationships by summarizing and accounting data that can guide further review & potentialize its outcomes.

4. RESULT & DISCUSSION
4.1 Top 30 Counties with Most Number of Geothermal/Oil/Gas Well
With the help of EDA, the graph shows the top 30 counties having the most number of oil/gas wells. The below graph shows Ellis county has most no. of wells because of which it significantly contributes to the world economy.

Following the finding of the producing commercial well on William Shutts' property in 1928, Ellis and the neighboring counties have been active in oil and gas production for almost forty years. From the late 1920s through the mid-1930s, large oil firms stepped inside and bought out small, independent oil companies, resulting in increased oil output. When oil and gas production resumed in the 1950s, oil prices fell, resulting in an uneven and less reliable oil market. Due to oversupply & drop in oil output and unpredictable markets, the oil corporations and enterprises that relocated to Hays were forced to close their doors or relocate. Despite shifting oil prices and the oil market's unpredictability over the years, the oil business should be acknowledged for the economic support is made on a personal, municipal, and state level.
4.2 Various Types of Wells in the USA
Exploratory Data Analysis (EDA) evaluates datasets to summarize their main characteristics, often using visual approaches. EDA is used before modeling to see what the data could tell us. With the help of it, we found 30 wells present in the USA.

It is found that D&A well is primarily present in the world whose count is 35642, the number of oil wells is 22808 which is more than gas wells, i.e., 19390. Brine well, junked well & pilot hole well are least compared to others based on data used for analysis.

Brines are an excellent source of salt (NaCl); during 500 BC, the Chinese drilled 100 brine wells, a few of which reached depths of more than 100 meters (330 ft.). Boreholes were used to dig large brine reserves beneath the Earth's surface. Surface water & groundwater are both included in commercial brines. Although brine produced by geothermal energy wells frequently contains large minerals, it is not currently utilized for industrial mineral extraction (Wikipedia, 2022). The study suggested better brine management practices to avoid negative environmental consequences and lower the economic disposal cost. The Zero Brine initiative attempts to achieve both of these goals while utilizing high-quality resources (Filtration & Separation, 2019).

Oil & gas are key commodities that are traded on a global scale. Transporting oil & gas products is relatively inexpensive and straightforward due to economies of scale. In many countries, oil and gas commerce contributes significantly (both positively and negatively) to their current account in the balance of trade. The number of oil exporters, in particular, rely mostly on exports for foreign cash (Stevens, 2018).

4.3 Number of Wells in Different States
According to finding, the most well present in Kansas State, i.e., 83293. Kansas' economy depends heavily on the petroleum sector. Since the late 19th century, hundreds or even thousands of natural gas & oil wells have been drilled throughout the state, yielding more than 6.7 billion barrels of oil & 41.2 trillion cubic feet of natural gas (Uwatoko, 1952). NE has 17199 wells and Texas has 8305 wells. Texas is the nation's leading producer of oil and natural gas. Texas will produce 43% of the country's crude oil and 26% of the nation's commercial natural gas in 2020. As of January 2020, Texas' 31 petroleum refineries could produce nearly 5.9 million barrels of crude oil per day, accounting for 31% of the nation's refining capacity (US Energy Information Administration, 2022).

Georgia and Northern California have 27 & 22 oils & gas wells, whereas G.A., Missouri, and S.D. have the least number of wells. Missouri has no major crude oil deposits and has produced roughly 100,000 barrels per year since the early 1980s, although output has declined since 2013. The state's yearly oil production was 75,000 barrels in 2020, down from a high of 285,000 barrels in 1984 (N Dakota, 2021). The below graph describes the actual count of geothermal, oil & gas wells in a different state.
4.4 Recorded Temperature Inside Well
Temperature measuring in wells is an old procedure, & geothermal gradients have long interested geologists' curiosity. Their use in oil/gas wells is a relatively recent development. Temperature abnormalities have long been noticed in drilling & producing wells, but thermometers capable of obtaining a reliable record of the abnormalities were unavailable (Millikan, 1941). The below graph shows the well temperature variation across the USA.

**Figure 4: Number of wells in a different state**

4.5 A Map-Based Plot of Wells Present Different Location
It is shown using the marker cluster function. The marker clustering function assists in managing several zones at various zoom settings. When a user zooms in to a high level, individual markers emerge on the map, which helps the process simple to examine the map. The
below graph shows the different mark points of wells present in other regions. The below map is the located map of the United States region.

Figure 6: oil/gas wells in U.S. region

In 2020, the total sum of dry natural gas manufactured in the United States was over 33.5 trillion cubic feet (Tcf), an average of 91.5 billion cubic feet per day, and the second-highest annual volume ever recorded. Horizontal drilling & hydraulic fracturing techniques have accounted for most production gains since 2005. In 2020, dry natural gas production in the U.S. was around 10% larger than total natural gas consumption in the United States (U.S. Energy Information Administration, 2022). Some of the oil/gas wells near Chicago are also shown below.

Figure 7: Wells in a different region of Chicago
4.6 Heatmap of Wells across the USA
A heatmap would be a graphical depiction of individual matrix values as colors. A heatmap is an excellent tool for displaying the intensity of values between 2 dimensions of a matrix. It helps with pattern recognition and provides a sense of depth. With its help, we represent the number of well types in different states see the below graph.

![Heatmap presentation of well distribution across the USA.](image)

**Figure 8: Heat Map presentation of well distribution across the USA.**

4.7 Prediction of Geothermal Well Temperature Using Machine Learning Algorithm
High-pressure, high-temperature prediction of wells is one of the most challenging circumstances in wellbore management because they need huge expenditures and maintenance expenses in the petroleum industry. In this context, drilling fluid rheology and its key factors will help engineers develop a fundamental knowledge of the best wellbore management strategy. Furthermore, it can help engineers in controlling the fluid loss issue. The authors have used two machine-learning algorithms to predict the geothermal surface temperature.

4.7.1 Support Vector Regression
SVR applies the same idea as Support vector machine but have regression issues. Now we will see working of implemented SVR to predict temperature.
The Scikit-Learn module in Python contains all of the functions required to implement SVR. All we have to do now is prepare a data set to train an SVR model.

- Import all essential libraries.
- Import the dataset and create the feature matrix & dependent variable vector.
- Feature scale the data.
- Lastly, fitting the SVR model to the dataset and predicting the result.
The RMSE of an SVR algorithm determines the model's absolute fit with the data. In other words, it displays how near the real data values are to the expected values of the model. A low RMSE number implies a better fit and is an excellent metric for measuring the reliability of the model's predictions.

### 4.7.2 XGBoost Regression

XGBoost is a decision-tree-based gradient boosting ensemble Machine Learning algorithm. (Solanki et al., 2021) Artificial neural networks (ANN) surpass all the frameworks or other algorithms in unstructured data prediction problems. For predicting a numerical value, like temperature, depth is the goal of regression predictive modeling issues (Brownlee, 2021). For predictive regression modeling, XGBoost can be utilized directly. By determining the root mean square error value for the XGB algorithm. It can be stated that RMSE values around 0.2 - 0.5 indicate that the algorithm can adequately predict the data.

The following steps were followed to implement the XGBoost model:

1. Import necessary libraries.
2. Set up parameters.
3. Predictors & target variables are being developed.
4. Divide the data into two groups: training and testing.
5. Initialize the XGBoost model.
6. Train the model.
7. Prediction evaluation.

#### Figure 10: Working of XGBoost Model (Ma et al., 2020).

The standard deviation of the residuals is defined as RMSE. Residuals measure the distance between the data points and the regression line; RMSE estimates all these residuals scattered. In other words, it indicates how dense the data is near the line of greatest fit. Root mean square error is often utilized in meteorology, forecasting, & regression analysis. There is no appropriate MSE value. Simply, the lower values, the better the model. As there is no true answer, the MSE's primary utility is deciding one prediction model over another. The equation below depict the mathematical notion of RMSE; where $N$ is the number of terms, $i$ is the $i^{th}$ Term. And the table 1 shows the RMSE values for the implemented model for their comparison.

$$ F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}) $$

where $\alpha_i$ and $r_i$ are the regularization parameters and residuals computed with the $i^{th}$ tree respectively, and $h_i$ is a function that is trained to predict residuals, $r_i$ using $X$ for the $i^{th}$ tree. To compute $\alpha_i$, we use the residuals computed, $r_i$ and compute the following: $\text{arg min}_{\alpha} \sum_{i=1}^{n} L(Y_i, F_{i-1}(X_i) + \alpha h_i(X_i, r_{i-1}))$ where $L(Y, F(X))$ is a differentiable loss function.
\[ RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\text{Predicted}_i - \text{Actual}_i)^2}{N}} \]  

(1)

Table 1: RMSE values for the implemented machine learning models.

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Machine Learning Model</th>
<th>RMSE Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>SVR</td>
<td>4.75</td>
</tr>
<tr>
<td>2.</td>
<td>XGBoost</td>
<td>5.16</td>
</tr>
</tbody>
</table>

5. CONCLUSION

Our study led us to conclude that EDA is a thorough examination that helps to find data structures. It is important for all industries because it reveals trends, patterns, and linkages that are not abundantly clear. EDA is the best method for identifying abnormalities, but it might mislead us wrong if not done correctly. Based on our analysis, it is concluded that:

- Most of the oil/gas wells, i.e., more than 7000 wells, are found in Ellis County, which significantly contributes to the economy.
- Many nations’ oil and gas commerce contributes significantly (both positively and negatively) to their current account in the balance of trade. The number of oil exporters, in particular, rely mostly on exports for foreign cash.
- According finding the most number of well present in Kansas State, i.e., 83293. Kansas' economy depends heavily on the petroleum sector, whereas Missouri has no major crude oil deposits and has produced roughly 100,000 barrels per year since the early 1980s.
- The usual geothermal gradient is thought to be roughly 1 degree Fahrenheit for every 60 feet of depth.
- We have used two machine learning models: Support Vector Regression and XGBoost regression, to determine which models better predict the geothermal surface temperature and determined the root mean square error value for both models, as shown in table 1.

Hence SVR's RMSE, i.e., 4.75, is less than XGBoost's root mean square error, 5.16. As a result, when compared to the XGBoost algorithm, SVR performs well in forecasting temperature utilizing independent factors.

REFERENCES


