This report focuses on the following challenge: Chevron wants to invest in renewable energy businesses and has tasked me, the data scientist, to identify the most promising US states for investment. My objective is to accurately predict the Renewable Energy Investments ($) of each US state for 2020.

To meet this objective, I plan to develop a predictive model that considers various factors that could impact renewable energy investments in a state. To ensure its accuracy, I will preprocess the data, encode categorical variables, optimize the model's hyperparameters and select appropriate algorithms. I will also inspect the features and their relevance to the response variable. It's worth noting that the data given is a time-series data distributed annually from 2015 to 2019, and my goal is to achieve the lowest RMSE for 2020 while ensuring the model is learning. This can be done by using other evaluation metrics, such as R2 score and MAPE and MAE.

I kickstarted the project with a data analysis and generated a pandas profiling report. The dataset had two challenges - missing values and categorical features. To handle missing values, I first explored the feature's distribution and found that the mean was higher than median. I grouped the data by state and replaced the missing values with the local median value. I also tried dropping the rows, but changes weren't significant, so I kept it simple by dropping rows since they only construed 1.3% of the complete data. For categorical variables, I tried one-hot-encoding but it resulted in high-dimensionality, which I reduced with PCA. The correlation matrix (as shown in Figure 1) showed that MSN and other features had little to no correlation with the predictor variable. I eventually used ordinal encoding as it didn't affect the models.

Visualizations showed that the data was balanced, but the response variable had many outliers due to its narrow distribution. I dropped the columns "Amount", "Index", and "State" as they didn't seem useful. Upon inspecting the MSN features it can be inferred that the states already using renewable energy are more likely to grow into it and thus require more investment compared to the states where traditional energy sources are cheap and abundant. Those states which are running out of the natural resources, possibly resulting in higher prices, might also be looking to move to renewables thus looking for investments. However, I don't have the expertise to engineer features based on this assumption.

I learned that a reliable approach to preprocess time-series data is to deseasonalize and then transform using a log transformation, but I discovered it too late for this project. I could only repeat the experiment with deseasonalizing.

I started with basic ML models such as Linear Regression, KNN Regression, Regression Trees and Random Forest using GridSearch. Its interesting to see that linear regression actually does better than all others. The results led me to conclude that a Random Forest Regressor could be a potential model for this task as it can handle a variety of input data types and is commonly used for time series forecasting. To improve its performance, I conducted a hyperparameter search using RandomSearchCV to find the best parameters.

Based on a brief research I found other models that could work well with forecasting problems such as MLP, Autoencoders, LSTMs, Encoders using LSTM so I tried using them and even fine tuning it but due to the lack of expertise, I was not able to achieve presentable results.

The analysis conducted revealed valuable insights into the relationships between various factors and target variable. Utilizing advanced statistical techniques and machine learning algorithms, I was able to build and experiment with various models.

The results of this study have implications for further research and practical applications in the field. I believe that the findings and methodology presented here could be used as a starting point for more in-depth investigations, potentially leading to improved decision making and outcomes.

Overall, participating in this datathon competition was a rewarding experience, allowing me to showcase my analytical and technical skills while solving a real-world problem.

Visualizations to achieve decisions about data are given below.

Diagram

Description automatically generated

Chart, histogram

Description automatically generated A picture containing chart

Description automatically generated Chart, histogram

Description automatically generated Chart

Description automatically generated with low confidence

Learning curve for Random Forest

Chart, line chart

Description automatically generated