NAME:PAVITRA BHARGAVA, ADM NO: 20JE0674

PREDICTING COAL PRODUCTION USING MACHINE LEARNING MODEL

INTRODUCTION

Coal has been a significant source of energy for every country for a very long time. Even in the United States, before the advent of renewable sources of energy, coal has been used as an important source of energy, providing a reliable as well as affordable fuel source for power generation and industrial processes. Coal has been mined in the USA through both surface and underground mining and any kind of disruption in its production can affect a country severely.

The purpose of this research paper is to develop a multivariate regression model that predicts coal production based on certain parameters. The model will examine the impact of three key variables on coal production: Average Employees, Labor Hours, and Mine Type. By understanding the influence of these factors on coal production, policymakers, and industry stakeholders can develop strategies to ensure the sustainability and stability of the industry and to support economic growth in coal mining regions. The results of this study will provide valuable deep insights into the dynamics of the coal mining industry in the United States and increase coal production through informed policies and strategies.

OBJECTIVE OF WORK

My work aims to develop a multivariate regression model that can accurately predict coal production in the United States based on publicly available production data. The model predicts coal production based on parameters “Average Employees”, Labor Hours”, and “Mine Type”. The model will help find out the factors that have the greatest impact on coal production levels and enable us to better understand the dynamics of the coal mining industry in the United States.

METHODOLOGY

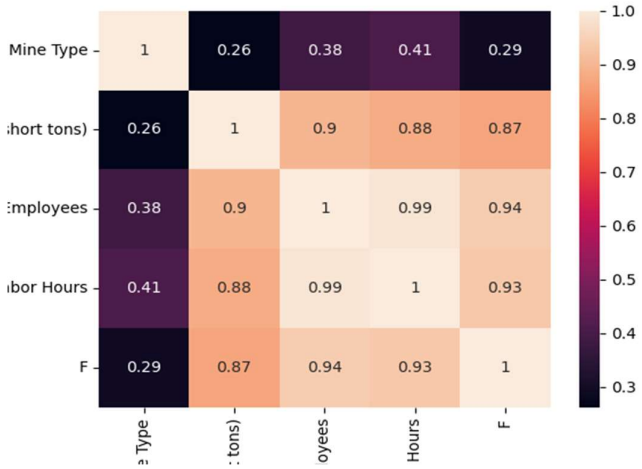
Data Collection

The data for this model was collected from the United States Government Site: Historical Coal Production Data: 2021 (Source: The U.S. Energy Information Administration (EIA) and the U.S. Mine Safety and Health Administration). The data provides comprehensive data on coal production in the United States. The data covers production data for the year 2020 and includes information on coal production, average employees, labor hours, and mine type for

each mine. The data was extracted from the website and imported into a Microsoft Excel spreadsheet for further analysis.

Data Analysis

The data for the model has been analyzed using a multivariate regression model. The dependent variable for the model was coal production, while the independent variables were average employees, labor hours, and mine type. An extra feature using the feature engineering model was also added to improve the model's accuracy, which was the multiplication of two of the parameters "Average Employees" and "Labor Hours". Before fitting the model, the data was preprocessed to ensure that it met the assumptions of the regression model. First, the data were checked for missing values, outliers, and extreme values. Missing values were replaced with the median of the corresponding variable, while outliers and extreme values were discarded.



Heat Map showing correlation matrix

Regression Analysis:

Once the data was preprocessed, a multivariate regression model was developed. The model equation used for regression is as follows:

Coal Production (short tons) = β0 + β1(Average Employees) + β2(Labor Hours) + β3(Mine Type)

The model was fitted to the data, and the coefficients (β0, β1, β2, and β3) were estimated to quantify the dependence of each of the input parameters on the dependent variable.

Model Training Evaluation

For Model Training and Evaluation, the standard 80-20 ratio for test data and train data was respectively used. To evaluate the model's accuracy, the R-squared value method was used. Mean absolute error, mean squared error and root mean squared error were also checked for. Additionally, residual plots were also plotted. The final model was used to

predict coal production levels for different scenarios, such as changes in average employees or mine type.

Results and Discussion

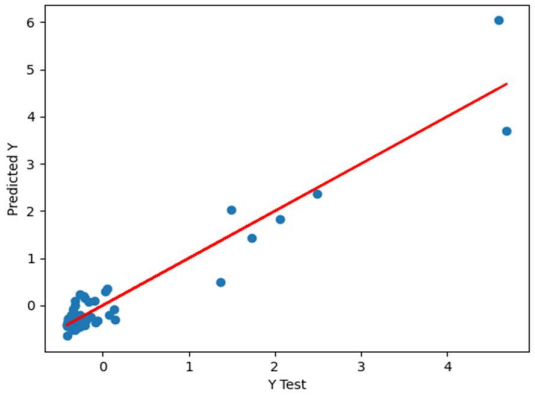
The coefficients of the models for the four parameters are shown below in image.

The R-squared value for the model came out to be 0.9047, indicating that the independent variables explain 90.47% of the variability in coal production. This suggests that the model is a good fit for the data and can be used to make moderately accurate predictions about future coal production levels.

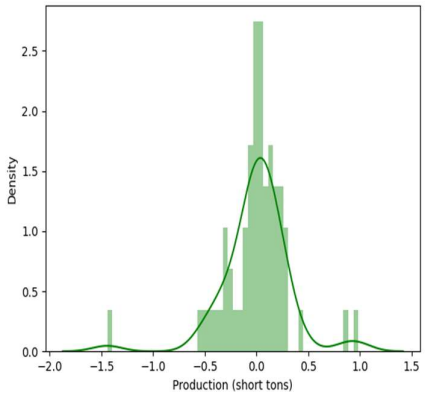
	Coeffecient
Average Employees	0.907769
Labor Hours	-0.165801
Mine Type	-0.221665
F	0.223502

The coefficient for average employees was positive, indicating that an increase in the number of employees is associated with an increase in coal production. The coefficient for labor hours was negative, indicating that an increase in labor hours above the optimum level is associated with a little decrease in coal production. The coefficient for average employees was particularly high, indicating that the number of average employees has a significant impact on coal production.

The model was also used to make predictions for different scenarios. For example, if the number of average employees, labor type, and mine type were taken as 5, 000, 20,000 and surface respectively, the model predicted that coal production would be approximately 22,351,463 short tons.



Predicted v/s Actual Values



Residual Plot

Conclusion

In this research paper, I developed a multivariate regression model to predict coal production in the United States based on average employees, labor hours, and mine type. The results of the model show that average employees have a significant positive impact on coal production, while labor hours have a negative impact on coal production. The model has an R-squared value of 0.9047, indicating that it explains 90.47% of the variability in coal production.

My work can have important implications for the coal mining industry in the United States. By understanding the factors that influence coal production, policymakers can formulate strategies to ensure the sustainability of the industry. For example, increasing the number of employees in a mine could lead to increased coal production.

However, it is important to note that the results of the model are based on historical data and are not foolproof and, may not necessarily hold true in the future. Also, there are certain other external factors that cannot be fed into the model.

References

1. United States Energy Information Administration. (2021). Coal Data Browser. Link: https://www.eia.gov/coal/data/browser/#/topic/29
2. Draper, N.R. & Smith, H. (2014). Applied Regression Analysis (3rd ed.). Hoboken, NJ: Wiley.
3. Hair, J.F., Black, W.C., Babin, B.J. & Anderson, R.E. (2010). Multivariate Data Analysis (7th ed.). Upper Saddle River, NJ: Pearson Education