JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 5123-5126
https://publishoa.com
ISSN: 1309-3452

Coal Production Estimation Using Regression Algorithms

Godavarthi Vinnu, Vannepaga Akhila, CH.Akhila Preethi
Under the guidance of
Mrs.B.Laxmi
JNTUH

Received 2022 March 15; **Revised** 2022 April 20; **Accepted** 2022 May 10.

Abstract

Background: As coal is providing major portion of energy requirements and it will continue to provide the same in the Future as well for the next few decades.It describes the information like amount, location, and quality of the coal resources. It is important that we need to use coal in an efficient manner as well as safely.We need to make sure that it is environment friendly so that environment is not effected by the things done by the human.As we can see that coal resources are mostly used in electricity. In this project we are going to estimate the coal production and to predict how much coal we have to use and how much we have to save for future purposes.So,we are using regression algorithms like Lasso and Ridge to estimate the coal production and we are going to compare both algorithms, choose the best one and fine tune the parameters of the algorithm.This project is beneficial for the coal production companies.

Methods:

1.Data Analyst:

We gather and interpret the data in order solve the problem we faced in existing system.

2.Model Implementation:

Here, we perform model training and model testing using advanced regression models.

3.Data Comparison:

we compare the predicted data with previous data in data comparison.

4.Data Visualisation:

In this we will see the results and visualize it.

Conclusions: While making a model most important things that we need to remember is feature engineering and selection process.To make the model more robust and accurate make sure to extract the maximum information from the features.The experience and time matters the most in feature selection.There are many ways to deal with our dataset and also many ways to make our model learn. The algorithms we are using should give better results compared to the existing system.For that we can use different algorithms and essemble them or compare all the algorithms and check which one is giving the best and accurate results.So, that it will be easy to choose which algorithm is perfect for our project.That model is used on our project and we can see the accurate results which will be useful for the coal based companies.

Keywords: Coal,Machine Learning,Resources,Lasso Regression,Ridge Regression,EnvironmentFriendly.

1. Introduction

Coal is the main resource for electricity, it is an abundant natural resource that can be used as a source of energy.The project main goal is to predict the coal production using parameters like no.of labours, working hours, area etc.We need to apply the best model to get the accurate information on coal.The dataset we are taking in this project is from Kaggle.The dataset has transactions of historical production of coal.we have to encode this data to discover the result.We are using linear regression models and these models would be trained.First we will take the half dataset for testing and other half for training.We will check whether the results are matching or not .Likewise we can also check which algorithm gives the accurate result.Coal has variable amounts of other elements,chiefly hydrogen,nitrogen,sulfur and oxygen.Coal is mostly used as a fuel.It has been used for many years and its usage was limited until the industrial revolution.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 5123-5126
https://publishoa.com
ISSN: 1309-3452

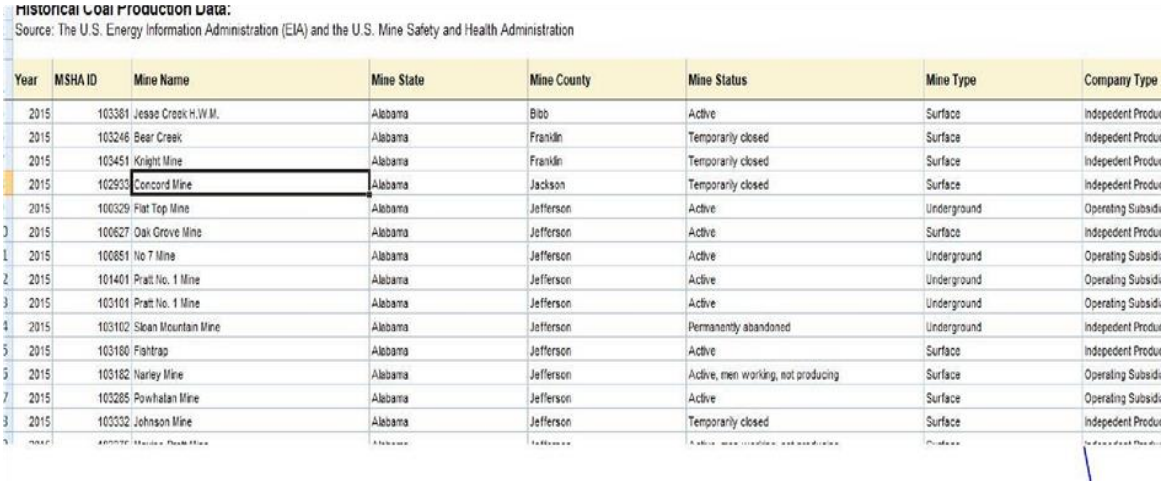
2. Objectives

Coal plays an important role in every one's life.As we are doing project on the coal we need to get full information about coal.So, in our project we are going to predict the coal production but using the linear regression we can't get the accurate results because it is not taking the full data and if we don't take full data then how can we get accurate results.It also suffers from overfitting.To overcome this all problem we are using another regression algorithms. The Existing system is using linear regression due to which we are facing many disadvantages like using this basic linear regression ,it does not have a control on your fitting parameters.Another one is overfitting data, does not provide enough preprocessing and also visualization.It also have EDA (Exploratory Data Analysis) issues. These are the few limitations that we face with the present system.It does not deal with the complex data. It suffers from overfitting because of no generalization of data. Due to this overfitting the errors on test data is high Our project main aim is to use advanced regression algorithms like ridge and lasso regression.First we take half of the data and train that data with this models and remaining data is used for testing. At last we compare and see the results of both training and testing data , if the results are matching then our project is successful and also accurate.Like this we do the project.The advantages of using these models are, it will understand and set the lower and upper bounds, it also deal with complex data , there is no chance of overfitting problem because lasso and ridge will solve that problem. Here first we will do prepare what type of data to collect, next according to the preparation of data we will collect the data. After that on the data we collected we will perform data preprocessing.Here the data is segregated, next we will do testing and training on the dataset. Next we will apply the lasso and regression models. Now as a result it will predict the data and last step is data visualization where we see the results and visualize it.

3. Methods

We have Four components or modules involved in this project.
They are 1.Data Analyst , 2.Model Implementation, 3. Data Comparison , 4. Data Visualization.
1.Data Analyst:
We gather and interpret the data in order solve the problem we faced in existing system.
2.Model Implementation:
Here, we perform model training and model testing using advanced regression models.
3.Data Comparison:
we compare the predicted data with previous data in data comparison.
4.Data Visualisation:
In this we will see the results and visualize it.

4. Results

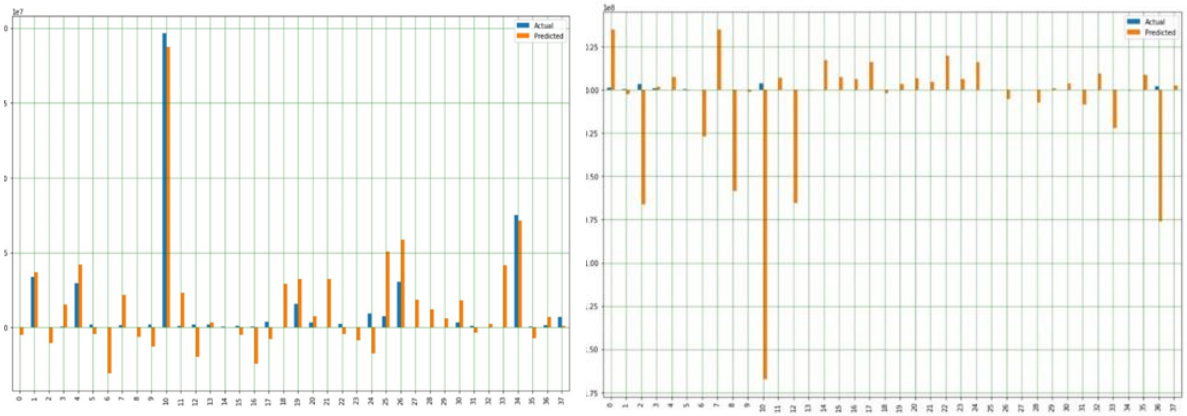


Historical Coal Production Data:
Source: The U.S. Energy Information Administration (EIA) and the U.S. Mine Safety and Health Administration

Year	MSHA ID	Mine Name	Mine State	Mine County	Mine Status	Mine Type	Company Type
2015	103381	Jesse Creek H.W.M.	Alabama	Bibb	Active	Surface	Indepedent Produ
2015	103246	Bear Creek	Alabama	Franklin	Temporarily closed	Surface	Indepedent Produ
2015	103451	Knight Mine	Alabama	Franklin	Temporarily closed	Surface	Indepedent Produ
2015	102933	Concord Mine	Alabama	Jackson	Temporarily closed	Surface	Indepedent Produ
2015	100329	Flat Top Mine	Alabama	Jefferson	Active	Underground	Operating Subsidi
2015	100627	Oak Grove Mine	Alabama	Jefferson	Active	Surface	Indepedent Produ
2015	100851	No 7 Mine	Alabama	Jefferson	Active	Underground	Operating Subsidi
2015	101401	Pratt No. 1 Mine	Alabama	Jefferson	Active	Underground	Operating Subsidi
2015	103101	Pratt No. 1 Mine	Alabama	Jefferson	Active	Underground	Operating Subsidi
2015	103102	Sloan Mountain Mine	Alabama	Jefferson	Permanently abandoned	Underground	Indepedent Produ
2015	103180	Fishtrap	Alabama	Jefferson	Active	Surface	Indepedent Produ
2015	103182	Narley Mine	Alabama	Jefferson	Active, men working, not producing	Surface	Operating Subsidi
2015	103285	Powhatan Mine	Alabama	Jefferson	Active	Surface	Operating Subsidi
2015	103332	Johnson Mine	Alabama	Jefferson	Temporarily closed	Surface	Indepedent Produ

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 5123-5126
https://publishoa.com
ISSN: 1309-3452

	Orginal	Prediction	Diff
0	2406437.00	2499572.75	-93135.75
1	1087329.00	1057502.88	29826.13
2	2155473.00	2117939.50	37533.50
3	194714.00	192289.36	2424.64
4	22011.00	22255.95	-244.95



5. Discussion

Ridge and lasso regression are simple techniques and also easy to use.They overcome the diadvantages we faced in existing system.

This dataset collected here is from United -states of America in the year 2015. This dataset consists of 1171 rows and 16 columns. The following steps are followed to solve the problem:

1.First Read the Libraries
2.Next Read the input dataset
3..From the dataset drop the unnecessary columns this is done by preprocessing
4.Wherever non-numerical values are present convert them to numeric values.
5.Also convert String values to float.
6.Divide the dataset into two that is training and testing.
7.Scaling the values of training and testing datasets
8.Last step is Fitting the Ridge and Lasso regression models

Refrences

1. https://colab.research.google.com/drive/1aTwWrzyU0uZvH0LkG4_zQxIJmoAPo1Lk?usp=sharing#scroll To=xezbOqVv98EM
2. https://www.kdnuggets.com /2018/02/essential-googlecolaboratory-tips-tricks.html.
3. Tolstoy Newton raja and AdityaNafde.(2011).Online Course Registration System Project Report. Florida: Computer and Information Science &Engineering at the University of Florida.
4. Responsive Web Design with HTML5 and CSS3 by Ben Frain, 2nd Edition
5. https://www.youtube.com/watch?v=49wBoO0bFMw&feature=youtu.be.
6. https://www.youtube.com/watch?v=SHYAQHDQoU4&feature=youtu.be.

7. Hazari, S., and Schnorr, D.(June1999). Leveraging Student Feedback to Improve Teaching in Web –based courses. T.H.E.Journals 26, no 11 30-32, 34, 36-38.(EJ589976)
8. Hedberg, J.G., and Corrent-Agostinho, S(June 2000): Creating a Postgraduate Virtual Community: Assessment Drives Learning. Educational Media International 37, no.2 83-90.
9. Hopper. (1998). Assessment in WWW Based Learning Systems: Opportunities and challenges. Journal of Universal Computer Science4, no. 4:329-347.
10. Marshall. (March 2000). Models Metaphors and measures: Issue in Distance Learning. Educational Media International 37, no. 1:2-8.