Quality Prediction of Iron Ore using Regression

Sai Vishal Tyagadurgam  
 Computer Science  
 Hood College  
Frederick, Maryland, USA  
st24@hood.edu

ABSTRACT

Regression analysis is one of the statistical process for estimating the relationships between a target variable and one or more predictor variables.

In this project I have created a prediction system that takes various parameters and predict the quality of Iron. I have applied different Regression analysis methods on the training data, using this I am able to predict the quality of iron.

CCS CONCEPTS

• Data Mining • Regression   • Data Analysis

KEYWORDS

Linear Regression, Random Forest Regression, Gradient Boost Regression, Ridge Regression

1 Introduction

Iron is generally extracted from rocks and minerals called as ores[1]. The quality of iron that is extracted is directly depended on the ore’s quality from which it is extracted. This quality is calculated by the percentage of silica that is present in it. The main aim of this project is to predict the amount of silica present in the iron ore.

2 Dataset

The dataset that is used in this project is a real time data that is collected during one of the several steps of a mining process of iron called Flotation Point. Most of the data is sampled at a regular interval of 20 seconds, some of it is recorded at an hourly basis. The data contains 24 columns[2].

The first column of each row is the date and time at which it is recorded. Second column tells about the percentage of iron feed present in the ore before it is introduced into the flotation plant. Percentage of silica present before feeding into flotation plant is recorded in the third column. The columns from 4 to 8 contain variables that impact the quality of ore like density, pH level, ammonia, flow of starch.

Once the ore is fed into the flotation plant there are variables that needs to be observed are recorded in the columns 9 to 22. In a lab the quality of an iron ore is calculated using the percentage of iron concentrate and the percentage of silica concentrate present in it. These 2 values are stored in the columns 23 and 24, respectively.

Chart, treemap chart

Description automatically generated

Figure 1: Correlation matrix

This is correlation matrix for the dataset showing how all attributes are related to one another. The highly correlated attributes are represented in lighter shades, darker shades indicate the attributes are weakly related to each other.

3 Methods

Regression is the data mining technique I have used in this project. Regression is the common technique used for problems like this where different attributes are used to predict a target. In this project to predict the target which is the percentage of silica present in the ore all the other variables are used. For this project I have used four different types of regression models.

3.1 Linear Regression

Linear regression is a linear approach in the field of statistics used to model a relationship between multiple variables[3]. In case of one predictor variable it is called as simple linear regression, for a more predictor variables it is multiple linear regression.

For this project linear regression is applied with all the default parameters. The default parameters include normalization being set to false, number of jobs set to none.

This model was unable to predict the quality, it was able achieve a success rate of 8.7% on both test set and training set.

Text

Description automatically generated

Figure 2: Linear Regression Results

3.2 Random Forest

Random forest is an ensemble learning method of regression which operate by fitting a few decision trees on various samples of the dataset and uses average to improve accuracy[4][5].

Unlike linear models, ensemble methods allow modification of several parameters to improve the performance. For this project I have worked with the depth of the decision trees ranging from the depths of 1 to 9.

Random forest regressor predicted the quality of iron ore with an accuracy ranging from 4% to 43% for the depths 1 to 9 respectively.

A picture containing graphical user interface

Description automatically generated

Figure 3: Random Forest Results

3.3 Gradient Boost

Gradient boost is a machine learning technique used for both regression and classification. It produces a predictive model from an ensemble of weak models[6].

Like random forest regressor, gradient boost allows to modify different parameters. I have worked with a different range of depths for the decision trees ranging from 1 to 9.

Gradient boost regressor predicted with an accuracy ranging from 13% to 74% for depths 1 to 9.

Graphical user interface, text, application

Description automatically generated

Figure 4: Gradient Boost Results

3.4 Ridge Regression

Like linear regression, ridge regression comes from the same family. This is particularly useful to mitigate the problem of multicollinearity in linear regression, which is a common problem seen in models with large number of parameters[7]. The main difference between these two regressors is the cost function. Ridge regression uses linear least square loss function.

Just like linear regression this is also performed without any changes to the default parameters.

Text

Description automatically generated

Figure 5: Ridge Regression Results

CONCLUSION

I hereby conclude that Regression can be used for problems like this where a single variable is predicted using the multiple variables. Among the four regression models I have used Gradient boost is the most efficient regression model with an accuracy of 74% at a depth of 9. Random forest was able to achieve an accuracy of 45% at depth 9. Both the linear models did not perform well for this data. Both the models produced an accuracy of 8%.

ACKNOWLEDGMENTS

I would like to thank Professor Dr. Xinlian Liu for his expert advice and guidance throughout the course as it helped me in learning various unknown topics related to data mining.

REFERENCES

[1] Ramanaidou and Wells, 2014

[2] Aditya Dehal (2019), Python notebook using data from Quality Prediction in a mining process <https://www.kaggle.com/aditya100/quality-prediction-in-iron-ore>

[3] [*David A. Freedman*](https://en.wikipedia.org/wiki/David_A._Freedman) (2009). Statistical Models: Theory and Practice. [*Cambridge University Press*](https://en.wikipedia.org/wiki/Cambridge_University_Press). p. 26. A simple regression equation has on the right-hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression e right hand side, each with its own slope coefficient

[4] Ho, Tin Kam (1995). [*Random Decision Forests*](https://web.archive.org/web/20160417030218/http:/ect.bell-labs.com/who/tkh/publications/papers/odt.pdf) (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from [the original](http://ect.bell-labs.com/who/tkh/publications/papers/odt.pdf) (PDF) on 17 April 2016. Retrieved 5 June 2016.

[5] Ho TK (1998). ["The Random Subspace Method for Constructing Decision Forests"](http://ect.bell-labs.com/who/tkh/publications/papers/df.pdf) (PDF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **20** (8): 832–844. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1109/34.709601](https://doi.org/10.1109%2F34.709601)

[6] Breiman, L. (June 1997). ["Arcing The Edge"](https://statistics.berkeley.edu/sites/default/files/tech-reports/486.pdf) (PDF). *Technical Report 486*. Statistics Department, University of California, Berkeley.

[7] [Kennedy, Peter](https://en.wikipedia.org/wiki/Peter_Kennedy_(economist)) (2003). [*A Guide to Econometrics*](https://books.google.com/books?id=B8I5SP69e4kC&pg=PA205) (Fifth ed.). Cambridge: The MIT Press. pp. 205–206. [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)) [0-262-61183-X](https://en.wikipedia.org/wiki/Special:BookSources/0-262-61183-X).