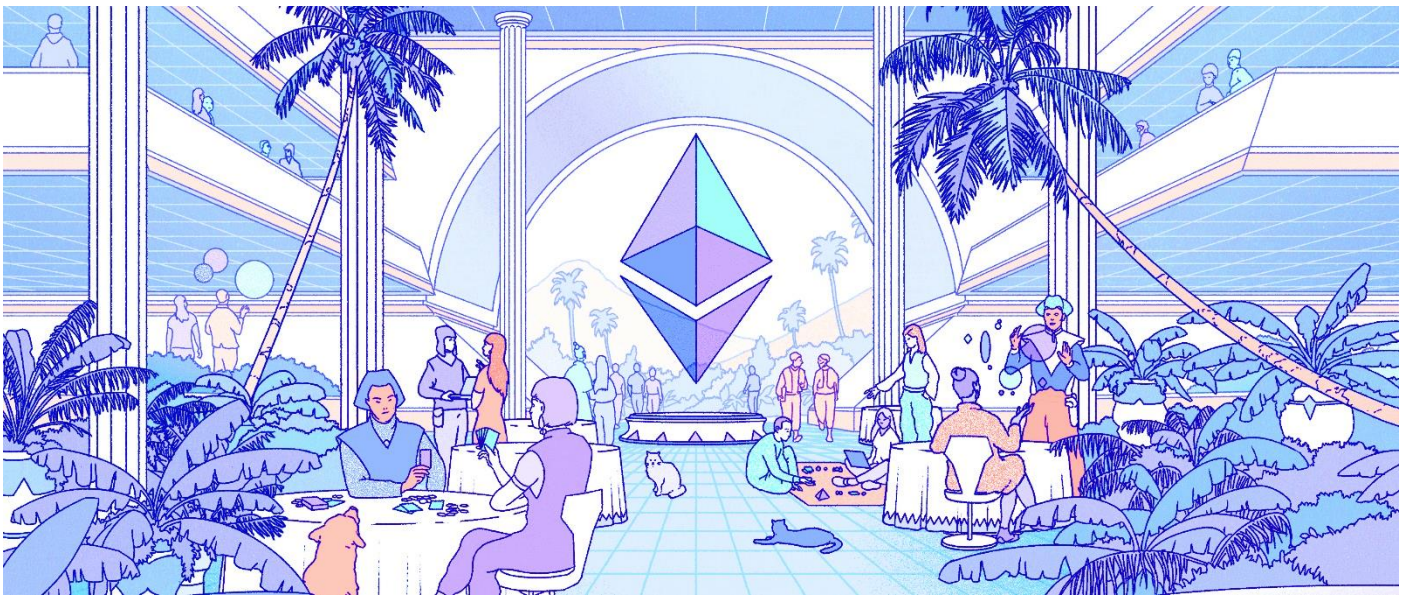
Ecological Mining on Ethereum

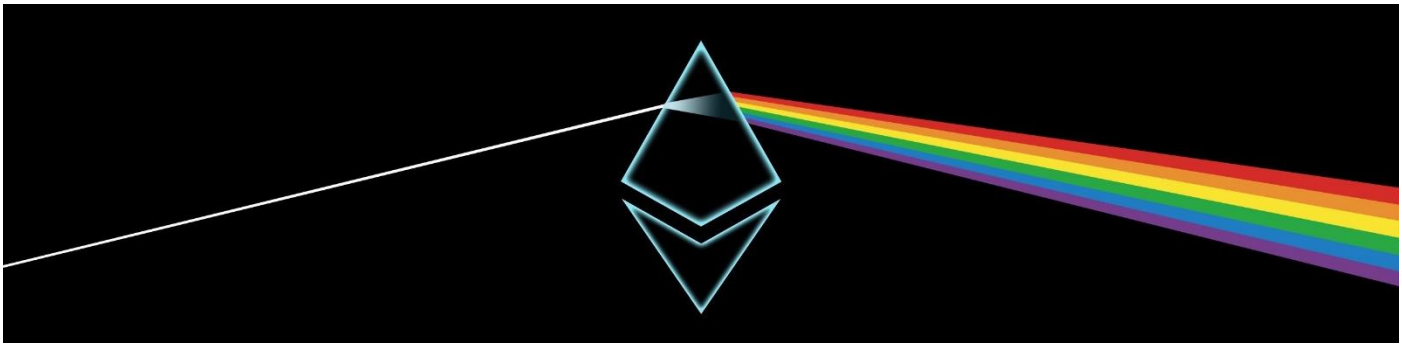


Group Members:

Uday Khokhariya

Saumya Patel

Riyank Makwana



Abstract

As cryptocurrency has gained popularity - and as more people understand how it works - its two most prominent players, Bitcoin and Ethereum, have come under the radar for the devastating environmental impact of the so-called mining operations that power their blockchains. According to Digiconomist's Ethereum Energy Consumption Index, these two networks alone consume more energy than the entire country of Thailand (population: 90 million). The recent surge in NFTs - digital art and other limited-edition collectibles stored on the blockchain - has focused attention on Ethereum, where the majority of NFTs are bought and sold. Currently, a single Ethereum transaction consumes the same amount of electricity as an average US household uses in a workweek - and has a carbon footprint equivalent to 140,893 Visa credit card transactions or 10,595 hours of YouTube viewing. Indeed, those multimillion-dollar NFT sales come at a high cost to the environment as well. One way of decreasing the environmental impact of Ethereum is to reduce the energy consumption for mining new blocks. Our model aims to predict the transaction fee required to mine a new block on the blockchain, so that next time when a new block needs to added to the chain, blockchain nodes will have an idea about the transaction fees which would be required. This would help in reducing overvaluation and overconsumption of energy. We are using the “On-Chain Ethereum Block Dataset” which can be obtained by querying the live Ethereum Blockchain. Since, the live network has more than 9,000,000 blocks, we would be taking a 0.1% data from a set of 1,000,000 blocks (11000000th to 11999999th block) for training while 200 for testing which would be completely different from testing data. We have used Multi Linear Regression Model (ML) which takes input parameters as size, difficulty, transaction count and average gas price, which in turn predicts the value of transaction fee required for mining a new block on the Ethereum Blockchain. Our model was able to decrease average loss from 556.857 to 16.9. We achieved the accuracy of 81.7% and 82.5% on training and testing data respectively. Since, the accuracies obtained from testing and training data are almost the same, our model is free from overfitting. We also incorporated individual predictor functions where user needs to set the parameters and he/she can obtain the predicted value of transaction fee based on them. This helps in real-life scenarios where the Ethereum Network needs an idea about the transaction fee and according to it, energy consumption of various nodes can be managed along with it. It also helps in distributing the fees among the blockchain nodes in a justified manner along with ensuring that nodes don’t utilize all the energy in the mining process. Thus, societal challenges which get addressed are “Energy Sustainability” and “Climate Action”. By reducing the blockchain mining, energy costs get reduced. This ultimately leads to bringing down energy requirements of the blockchain as well as saving the environment from harmful effects of Global Warming.

Index

S. No.	Chapter	Page No.
[1]	Introduction • *Main problem being targeted* • *Societal challenges being addressed*	4
[2]	Dataset • *Data description* • *Data source and discussion*	5
[3]	Architecture • *Block Diagram of the Model* • *Block Diagram Explanation*	6-7
[4]	Experiment • *Description of Steps performed* • *Can-View link of Google Colab*	8
[5]	Results • *Description of results obtained* • *Tables*	9
[6]	Conclusion • *Contributions* • *Achievements* • *Addressing Societal Challenge* • *Shortcomings & Future Scope*	10

1 Introduction

Cryptocurrencies have come a long way since their inauspicious beginnings. While the mainstream financial world once dismissed the usage of digital currencies, the industry has made significant progress towards establishing cryptocurrencies as a legitimate and world-changing space. With Bitcoin and Ethereum getting massive growth in price and users, there are still various effects that need to be considered for widespread adoption. Environmentalists have expressed grave concern regarding energy consumption of cryptocurrency mining which leads to increasing carbon emission and climate change.

Bitcoin receives the majority of the attention and scorn heaped on cryptocurrencies, leaving its younger sibling Ethereum in the shadows. However, Ethereum is far from insignificant. Its market capitalization has grown more than US $10 billion and it has an equally massive energy footprint. Bitcoin and Ethereum both work on proof-of-work consensus protocol which requires huge amount of computational efforts and vast energy consumption. Ethereum mining consumes a quarter to half the energy that Bitcoin mining does, but it still consumed roughly the same amount of electricity as Iceland for the majority of 2018. Furthermore, the average Ethereum transaction consumes more power than the average US household consumes in a day.

"That's just a huge waste of resources, even if you don't believe that pollution and carbon dioxide are an issue. There are real consumers—real people—whose need for electricity is being displaced by this stuff."
-Vitalik Buterin
(The 24-year-old Russian-Canadian computer scientist who invented Ethereum at the age of 18.)

Ethereum is built on a blockchain, which is a digital ledger of transactions maintained by a user community. (It's called a blockchain because new transactions are bundled into "blocks" of data and written onto the end of an existing "chain" of blocks that describe all previous transactions.) Hence, Ethereum doesn't require any central authority. It can be imagined as a global computer that is decentralized, open to all, and virtually immune to downtime, censorship, and fraud. The Ethereum blockchain's ability to store data, support decisions, and automate the distribution of value is what gives it such potential. These tasks are managed by smart contracts, which are programs written by users or developers in Ethereum's custom coding language. Smart contracts have obvious business applications, but the long-term hope is that apps built on them will make Ethereum the ultimate cloud-computing platform.

The issue is all of the mining. Ethereum, like most cryptocurrencies, is based on a computational competition known as proof of work (PoW). Because of the competitive nature of proof-of-work blockchains, these astronomical energy costs exist. Instead of storing account balances in a central database, cryptocurrency transactions are recorded by a distributed network of miners who are rewarded for their efforts through block rewards. These specialised computers are in a race to record new blocks, which can only be created by solving cryptographic puzzles.

"The Ethereum network consumes as much power as the entire nation of Qatar."
-According to Digiconomist, a cryptocurrency analytics site.

Coal and other fossil fuels are currently a major source of electricity around the world, powering cryptocurrency mining operations as well as other industries. However, the carbon dioxide produced by the coal-burning process contributes significantly to climate change.

One possible way of reducing the energy consumption is to significantly reduce the overestimation of gas fees required to make a new block on the blockchain. In order to do that, we need a rough idea of transaction fees which would be required in order to mine a block successfully on the decentralized network. Hence, we would be predicting the transaction fees required to mine a block successfully on the Ethereum Blockchain. We would be utilising the parameters of block size, difficulty, number of transactions and average gas price to make a model which could efficiently predict what should be required transaction fee in order to mine a new block. Mining rewards are distributed according to the transaction fees. Due to fluctuating nature of the Ether, transaction fees required to successfully complete a transaction also varies. So, next time whenever we would have to mine a block, transaction initiator would be having an average ether price that would be required in order to proceed for a successful validation. Therefore, the societal challenges which we would be addressing will be "Energy Sustainability" and "Climate Action".

Moreover, the results can be further expanded by making the data open-source to the miners and let the blockchain network decide the energy consumption based on predicted transaction fees. This would allow saving energy for the next block as well as prevent overconsumption of electricity. According to CNBC, cryptocurrency mining causes the release of 35.95 million tonnes of carbon dioxide into the environment, which in turn leads to an increase in the greenhouse effect and, eventually, global warming. Thus, forecasting transaction fees would aid in preventing the overvaluation of gas fees and, in addition, could be used to limit miners on the Ethereum Blockchain. This would aid in reducing excessive power consumption and, as a result, in saving the environment from Global Warming.

2 Dataset

We would be using the "On-Chain Ethereum Block Dataset" which could be obtained by querying the live Ethereum Blockchain. Since the full node of Ethereum has more than 9,000,000 blocks, we would be only be taking a part of that dataset for the project which in future, can be expanded and applied to all the blocks. We have fetched the required dataset from xblock.pro website (http://xblock.pro/tx/). Xblock.pro runs a full node (up to 13,249,999 blocks) and records the on-chain transaction data. We would be running our model for a part of the dataset consisting of 1,000,000 blocks (11000000th to 11999999th block). The dataset name is "Ethereum On-chain Data".

Raw Data Stats:

- Total Rows: 1,000,000
- Total Columns: 17

Columns of the raw data are:

- blockNumber – block number of the stored block on the blockchain
- timestamp – timestamp at which the block was created
- size – block size
- difficulty – difficulty of the blockchain at the time of creation of the block
- transactionCount – total number of transactions stored in the block
- internalTxCntSimple – total number of simple transactions stored in the block
- internalTxCntAdvanced – total number of advanced transactions stored in the block
- erc20TxCnt – erc 20 transaction count of the block
- erc721TxCnt – erc 721 transaction count of the block
- minerAddress – miner address of the block
- minerExtra – miner extra of the block
- gasLimit – gas limit associated with the block
- gasUsed – gas used for the creation of the block
- minGasPrice – min gas price at the time of creation of the block
- maxGasPrice – max gas price at the time of creation of the block
- avgGasPrice – avg gas price at the time of creation of the block
- txFee – transaction fees required to mine the block

Dataset taken for the model:

- ✓ Train Data Records: 1,000 (Randomized)
- ✓ Test Data Records: 200 (Randomized)
- ✓ Columns: 5 (size, difficulty, transactionCount, avgGasPrice, txFee)
- ✓ Features (Predictor Variables): 4 (size, difficulty, transactionCount, avgGasPrice)
- ✓ Response Variable: txFee

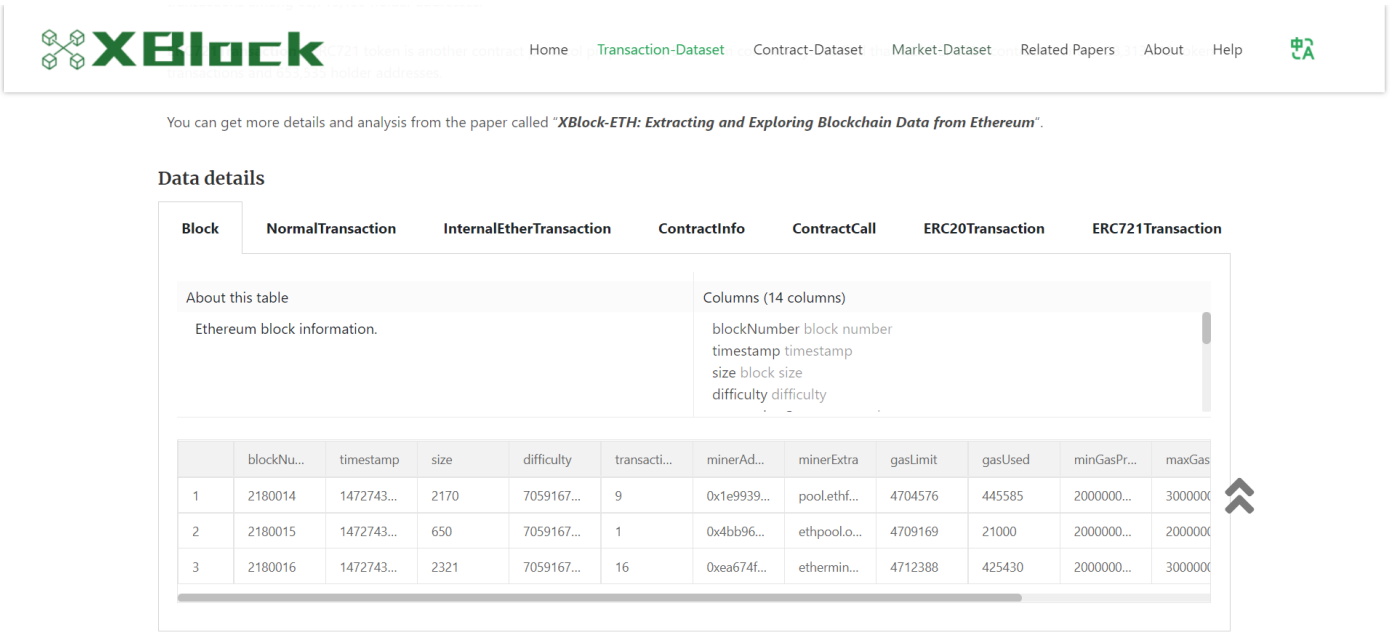


Fig 1. Screenshot of Xblock.pro, dataset source site.

3 Architecture

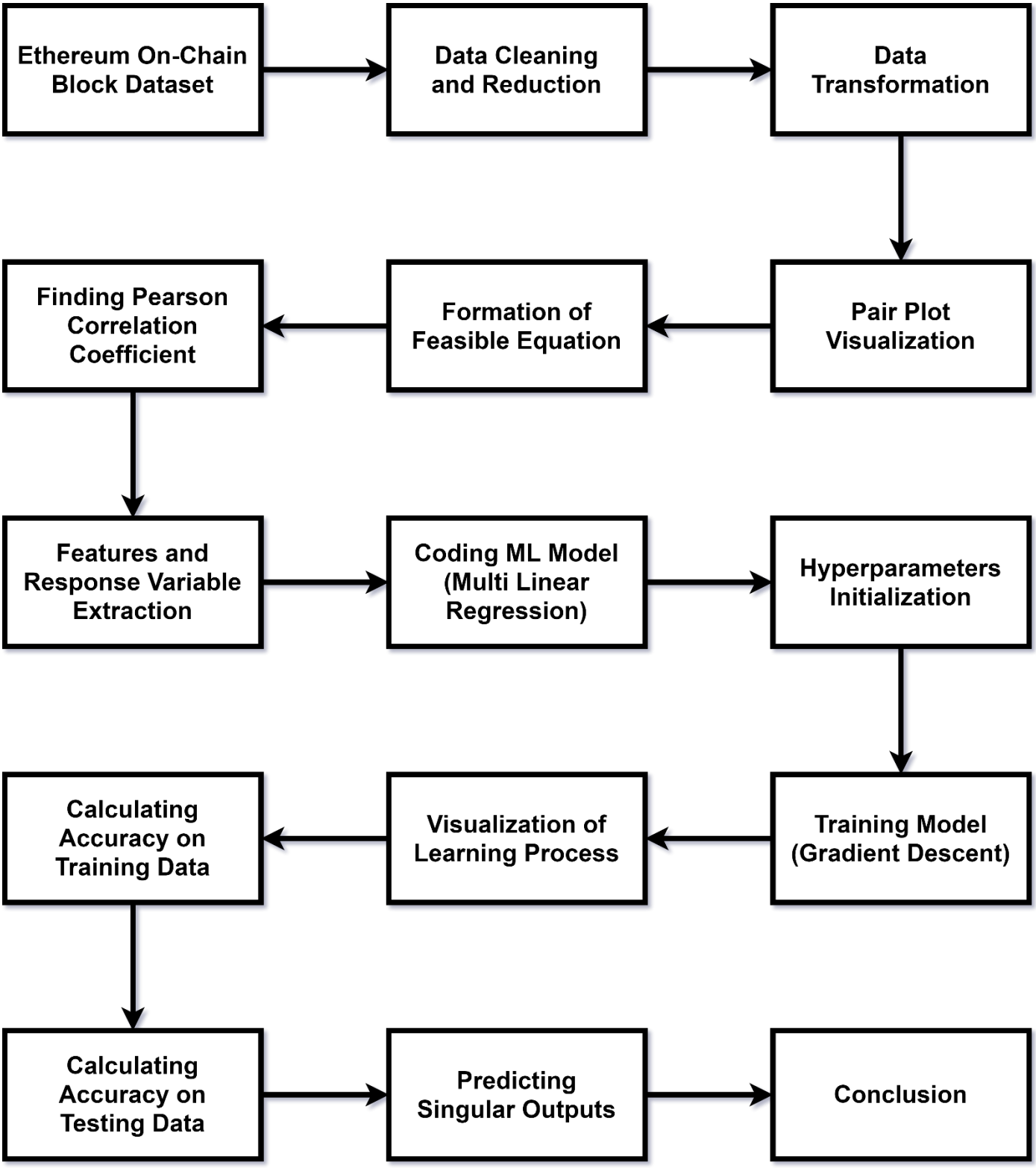


Fig 2. Block Diagram of the Project.

Description of the Block Diagram:

- ***Ethereum On-Chain Block Dataset:*** "On-Chain Ethereum Block Dataset" consisting of 1,000,000 blocks (11000000^{th} to 11999999^{th} block) is obtained from xblock.pro website which records the data of the live Ethereum Network.
- ***Data Cleaning and Reduction:*** Since, the raw dataset is huge, we would be taking 0.1% part of it (1,000 records) for training the model and another 200 for testing the model. Raw data had 17 columns, we would be taking only 5 useful columns (4 features and 1 response variable).
- ***Data Transformation:*** Since, the original values are too huge to be calculated and stored, we would be transforming the data records for faster processing. We found min and max values of each column in train dataset and scaled all values of columns using Min-Max Scalar.
- ***Pair Plot Visualization:*** We visualized the train dataset through pair-plot visualization between pair of all variables in the dataset.
- ***Formation of Feasible Equation:*** Through graphs obtained between different variables, we formed an equation for the model.
- ***Finding Pearson Correlation Coefficient:*** Pearson Correlation Coefficient is found between each pair of variables to get an idea about their dependency on response variable (transaction fee).
- ***Features and Response Variable Extraction:*** 4 Features (size, difficulty, transaction count, average gas price) and 1 response variable (transaction fee) are extracted.
- ***Coding ML Model (Multi Linear Regression):*** Multi Linear Regression model class is coded which would make the predictions of the transaction fee based on required input parameters.
- ***Hyperparameters Initialization:*** Hyperparameters like Epoch, Batch Size and Learning Rates are initialized for the model.
- ***Training Model (Gradient Descent):*** Model is trained using the train dataset.
- ***Visualization of Learning Process:*** Learning phase is visualized with the help of graphs in order to get an idea about what happened during the training of the model.
- ***Calculating Accuracy on Training Data:*** Accuracy on training data along with correct and incorrect results are obtained.
- ***Calculating Accuracy on Testing Data:*** Accuracy on testing data along with correct and incorrect results are obtained.
- ***Predicting Singular Outputs:*** Model function is called to predict the transaction fee based on given input parameters.
- ***Conclusion:*** Our Multi Linear Regression model is ready and can be used for further predictions as well as can be applied to numerous blocks of the Ethereum Blockchain.

Learning Model:

- ***INPUT:*** The input parameters are size, difficulty, transaction count and average gas price.
- ***OUTPUT:*** The output of the model is transaction fee.

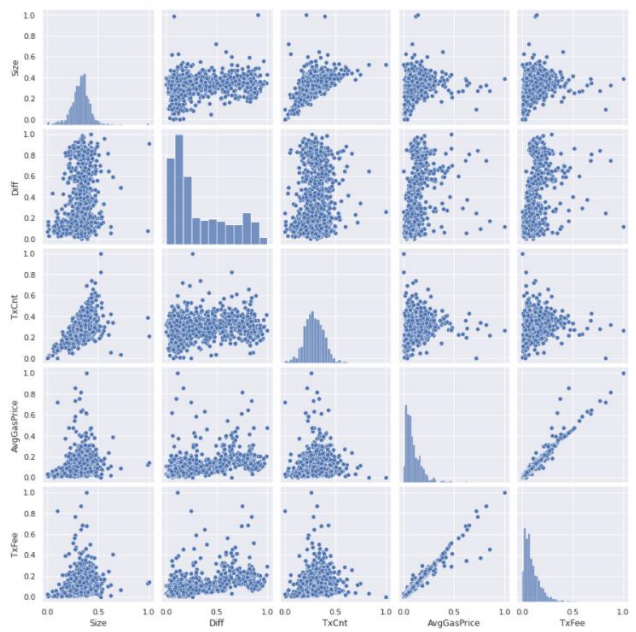


Fig 3. Pair plot between all pairs of variables.

4 Experiment

Steps involved in the development of the project are:

i. First, we have imported all the useful libraries, that Machine Learning Engineers use day-to-day in their programming career, like Pandas, Numpy, Math, Matplotlib and Seaborn.
ii. We have cleaned and reduced the dataset according to the requirements.
iii. Then, we have imported 2 datasets (train dataset and test dataset) into our Google Colab.
iv. We also then, looked at 5 datapoints of each dataset, to get an overview of how the data is.
v. We looked at the stats of each dataset.
vi. To increase readability of code for the rest of the project, we have renamed the big column names to comparatively smaller ones.
vii. To have a homogeneous datatype over all the columns, we have converted all the values of each column to float.
viii. Then, we have Summarised the dataset.
ix. We found Minimum and Maximum values of each column which would help in data transformation.
x. Range of values in all the columns is different, which can hinder the model from being optimum; so, we squeezed the values of each column within the range of 0 to 1.
xi. We displayed Pair-Plot to gather the insights related to the relation between the variables.
xii. We measured the Pearson-Coefficient of each pair of variables, in order to know the strength of their relationship.
xiii. We then plotted graphs (scatter) between all the 4 variables against TransactionFee (5^{th} variable).
xiv. Then, we extracted all the 4 predictors from datasets to a numpy array (X_train, X_test), and the dependant variable to another numpy array (Y_train, Y_test).
xv. We coded Multi Linear Regression Model (ML) class which can be used afterwards for making predictions.
xvi. Then, initialization of hyperparameters like Epochs, Batch size, Learning Rates, Weights, Bias is done.
xvii. Model is trained by calling the train function of the class and passing the hyperparameters to it.
xviii. The changes in the values of loss, gradients, learning rates, weights and bias during the training phase is plotted.
xix. Threshold value for Accuracy of Model on Training and Testing Data is been set.
xx. Accuracy of model on Training Data is calculated.
xxi. Model is tested using the Testing Data.
xxii. Model is then used to predict the output for any user-defined inputs.

Can View link of Google Colab:

https://colab.research.google.com/drive/1Xx8Ci1fSFhZK0-FQpcqwPdH0HowpILLb?usp=sharing

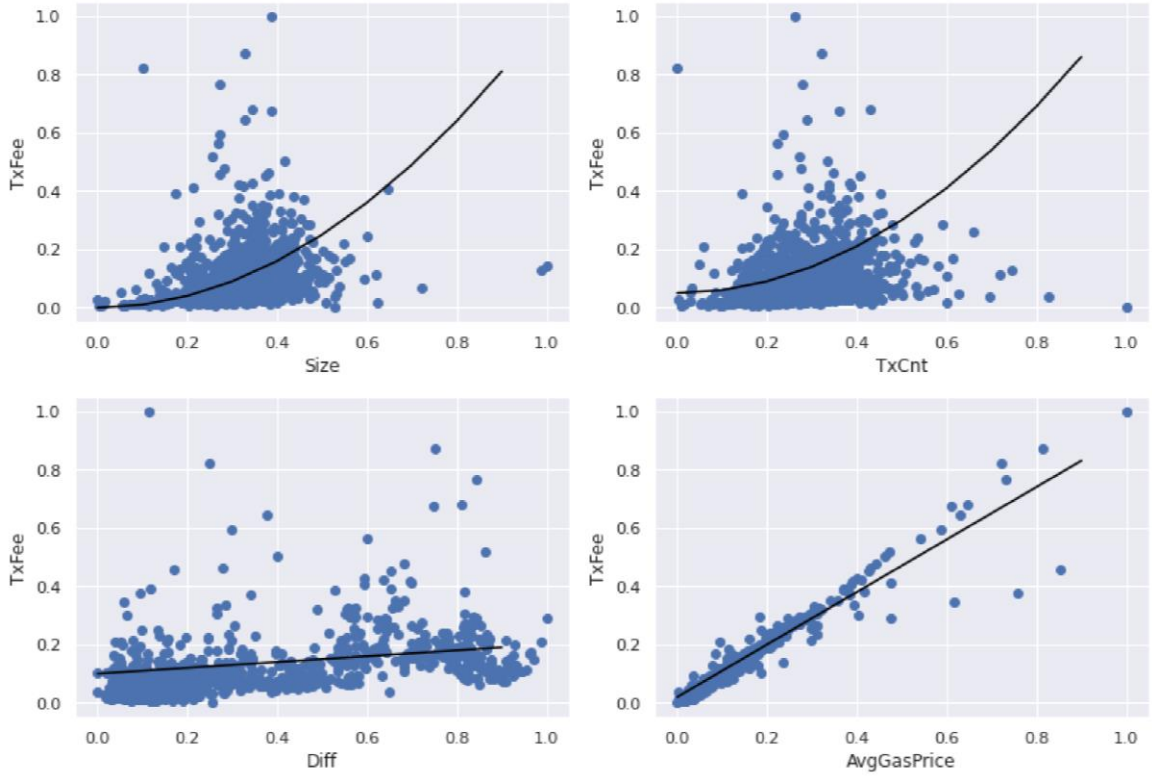


Fig 4. Speculating the trend.

5 Results

The results obtained from the implementation of Multi Linear Regression Model on "On-Chain Ethereum Block Dataset" are:

Final Equation:

*TxFee = w1 * Size^2 + w2 * Diff + w3 * TxCnt^2 + w4 * AvgGasPrice + bias*

HyperParameters & Parmeters:

- Size of Training Dataset: 1000
- Number of Epochs: 1500
- Batch Size: 50
- Initial Learning Rates (lr): [0.01 0.01 0.01 0.01 0.01]
- Initial Weights: [0.5, 0.5, 0.5, 0.5]
- Initial Bias: 0.5
- Initial Equation: TxFee = 0.5 * Size^2 + 0.5 * Difficulty + 0.5 * TxCnt^2 + 0.5 * AvgGasPrice + 0.5

Model Training:

- Initial Avg Loss: 556.857
- Final Learning Rates (lr): [0.00000000e+00 0.00000000e+00 0.00000000e+00 6.10351563e-07
- 0.00000000e+00]
- Final Avg Loss: 16.9
- Final Equation: TxFee = 0.417 * Size^2 + 0.122 * Difficulty + 0.429 * TxCnt^2 + 0.24 * AvgGasPrice + -0.067

Threshold of difference:

- Threshold: 0.07 (For Both Training and Testing Data)

The accuracies obtained in the implemented model are:

- *Training Data:*
  - Correct results: 817 / 1000
  - Accuracy: 81.7 %
- *Testing Data:*
  - Correct results: 165 / 200
  - Accuracy: 82.5 %

Predictions by the model, for user-defined inputs:

S. No.	Size	Diff	TxCnt	AvgGasPrice	TxFee (Predicted Output)
[1]	10000.0	3510013500000000.0	117.0	789001070000.0	1.917530669674259e+18
[2]	70413.0	5170012400000000.0	107.0	789135620000.0	4.146299596018587e+18
[3]	98932.0	4672013500086470.0	379.0	823911294000.0	7.320170196775355e+18

6 Conclusion

6.1 Contributions

We have formed the Multi Linear Regression Model which can successfully predict values of a new block to be mined by the nodes on the Ethereum Network on the basis of required parameters given as input. We formed the equation after the analysis of the graphs and pair-plot visualization, along with Pearson Correlation Coefficient between each pair of variables. We have also formed individual functions which can accomplish the task of finding answer to a single query. Visualization of the training phase is provided to get more insight into how the model actually works.

6.2 Achievements

We were able to bring down average loss from 556.857 to 16.9. Under training data, we were able to get 817 correct results out of 1000, while under testing, we got 165 correct results out of 200 with the threshold as 0.07. The accuracies obtained in train and test data were 81.7% and 82.5% respectively. As you can see that the accuracies, achieved by the ML model when training data and testing data were passed to it, are nearly the same; so, we can conclude that there is no case of overfitting in our model. Hence, the model is the best fit to the Ethereum dataset.

6.3 Addressing Societal Challenge

The results obtained from the model can be successfully applied to mine new blocks on the blockchain. Societal challenges which we are addressing are "Energy Sustainability" and "Climate Action". The exponential growth of Ethereum Cryptocurrency Market has led to increasing number of miners who compete with one another to mine the block first and earn a substantial mining reward in the process. This leads to vast utilisation of energy and power because every node on the Blockchain would be competing to mine the block first. Since, transaction fees are limited, it does not fairly compensate the energy utilised in the process. Thus, prediction of transaction fees would help in preventing overvaluation of gas fees and furthermore, can be used to limit the miners on the Ethereum Blockchain. This would help in reducing the exaggerated power usage and would consequently help in saving the environment from Global Warming.

6.4 Shortcomings & Future Scope

All the transactions happening on the Ethereum Blockchain are distinct and of varied complexity due to newer smart contracts getting added every day to the live network. These transactions require varying gas fees in order to proceed and successfully complete its desired task. Our model doesn't take into account the different types of transactions which are happening (ranging from simple to advanced) as it runs on total number of transactions which are happening in general. With the introduction of NFTs (Non-Fungible Tokens), there also comes a need of separating the types of tokens as well (like ERC-20 and ERC-721). We can get more precise values of transaction fees if we incorporate individual contributions of NFTs and distinguish between simple and advanced transactions. Moreover, individual contribution of those transactions in terms of gas would help in getting a more realistic idea about the results. Also, the average gas price depends on the market value which is often subjected to change due to changing market scenario. Thus, due to fluctuating nature of gas price, it becomes difficult for model to follow a particular trend in order to make the precise prediction. Currently, we have applied our model to a quite small amount of live chain data but, it can be further expanded to all the blocks on the blockchain to get more realistic idea about the predicted transaction fees.