# Analytics Startup Plan

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | *Prediction of Fraud Transaction in Ethereum* |
| **Requestor** | *Centennial College* |
| **Date of Request** | *15th July 2022* |
| **Target Quarter for Delivery** |  |
| **Epic Link(s)** | *Not applicable* |
| **Business Impact** | *To increase new and loyal users by building more reliable and compatible Ethereum platform* |

## 1.0 Business Opportunity Brief

|  |  |
| --- | --- |
|  | Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment. |

Ethereum

Ethereum is a kind of digital currency platform with its own value called ether. People can use digital currency such as ether, bitcoin, etc. as for the use in financial transaction in online games and advertisements, paying for goods and services, investing to trade into real money, and so on. Ethereum works in a decentralized block-chain system in which new nodes are always mined to keep the increasing transactions in Ethereum.

Business Problem

As the Ethereum becomes the second largest digital currency platform, there are more than 1 million transactions in a day in 2022. For the increasing number of transactions day by day, fraud transactions are mixed and can be found in the Ethereum. Some users are scammed and lost their value on Ethereum because of these fraudulent activities on Ethereum software.

**The specific ask:**

*Clearly articulate the specific task you will be conducting to help achieve the opportunity*

My goal here is to predict the fraud transaction before it can make any scams to real users and make an alert and be able to stop that transaction. First, the characteristic of fraud transactions will be classified and determined by building a model in python and by using the model, the Ethereum system will block the transaction which seems to be fraud depending on the selected characteristics such as history of user’s account, number of transactions, amount of ether, ether’s contract, etc.

## 1.1 Supporting Insights

|  |  |
| --- | --- |
|  | Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally? |

Ethereum.org offers a digital bank for users with its currency, ether which can be transferred between nodes of a blockchain system on ether digital platform for any use in applications. Each node contains accounts for users, smart contract code and smart contract state. All the information, transactions and changes made by users can be traced back on ether nodes.

Anyone around the world with an accessible internet can use Ethereum wallet to invest on digital currency, to pay for advertisements and applications, to buy online game items and services, to trade in with other digital currency or USD dollar.

The market share of Ethereum is around 120 M by June 2022. 1 ether equals to $ 1,113.50 USD by 13th July. When it was first launched as in August 2014, the value of ether costed only $0.31 USD. (CoinDesk, n.d.)

The key competitors of Ethereum are

* Bitcoin (the largest cryptocurrency platform)
* Tether
* USD coin
* Binance coin
* Solana
* Cardano
* Tezos
* Polkadot

(DeMatteo, 2022)

## 1.2 Project Gains

|  |  |
| --- | --- |
|  | *Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI’s.* |

From this project of predicting fraud transaction in Ethereum, the organization will get more trust from users and this Ethereum software will be used more widely around the world. Then, after increasing new users and having more transactions per day, the rate of one ether will be increasing day by day which will benefit both users and the Ethereum organization itself.

In this project, I will use the data set which includes information related to transactions in Ethereum with known fraud transaction. By using this data set, I will build a model to predict the fraud using the key factors from the data, then define the prevention and recommendation of how the organization should improve its infrastructure.

The use of this project will help blocking hackers and scammers to make fraud transactions on Ethereum digital platform and help the real users from losing their property on ether.

## *Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.*

## 2.0 Analytics Objective

|  |  |
| --- | --- |
|  | List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.  Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed. |

## The objectives of the analysis are shown as following:

* To find the key factors that are mostly related to a fraud transaction
* To choose the best model with the best accuracy to predict the fraud transaction
* To analyze alternative improvements to be done in the system

## 2.1 Other related questions and Assumptions:

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis* |

* The value of Ether will increase up to more than 300% nowadays as it increased up to 300% from $1500USD to $4500USD per ether in 2021. This changes of the rate of ether everyday may impact the analysis from this project.

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  | *What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.* |
|  | * The best model will be picked and predict the fraud transaction. * The activities of hackers and scammers on Ethereum can be significantly reduced. * Rate of Ether will be increased gradually. * Increasing new users, more transactions, and investments on Ether around the world. |
|  |  |

## 2.3 Methodology and Approach

|  |  |
| --- | --- |
|  | *Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.* |

**Type of Analysis:** *logistic regression, random forest regression, XG boost or KNN*

*First of all, the three or four models are used to predict the occurring of fraud transaction using the mostly related variables such as max, min, avg number of transactions per account, value received, sent of an account, etc. Then, compare the models and choose the best model that gives more precise result.*

**Methodology:** *Key questions from ‘Analytics objective’ will be tackled in ascending order as outlined in ‘5.0 Timelines and deliverable section’.*

*First, data cleaning and data visualization will be done if there are any missing values, outliers and multicollinearity. We will choose the most related variable to the objective of the project and remove unnecessary columns. We will use a heat map to find the correlation between variables to reduce multicolinearity and define the key factors affecting the fraud transaction. As there are some categorical variables in the dataset, those variables will be inverted into numerical values if necessary.*

*After that, as there are some columns with zero values which will not help anything in building a model, some dimension reduction techniques will be used to optimize the dataset to become more accurate and clean data. One thing to be aware is that the dataset is imbalanced so if there might be some problems because of it, we will find a way to cope it.*

*Then, we will set the response variable to be 1 if the transaction seems to be fraud and 0 if the transaction is normal. After that we build a model to predict the nature of transaction fraud -1 or not -0 by using the characteristic of selected variables with respect to the response variable.*

*In other words, we will split the data into train and valid data sets, and then build a logistic regression model and predict the model. Then, use the confusion matrix to see the accuracy of the model. We repeat the process with the other models and compare the accuracy between them.*

**Output:** *The output will be a set of insights, predictions, and recommendations to detect the fraud transaction to any user account or to the Ethereum platform with red flag and so that scammers and hackers can be blocked before they make a transaction.*

## 3.0 Population, Variable Selection, considerations

|  |  |
| --- | --- |
|  | Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today. |

**Audience/population selection: Ethereum platform users, professors, advisors**

**Observation window: 2020**

**Inclusions:** *all variables except exclusion and some which are very correlated.*

**Exclusions:** *Index, Address*

**Data Sources:** <https://www.kaggle.com/datasets/vagifa/ethereum-frauddetection-dataset>

**Audience Level:** Technical team of Ethereum

**Variable Selection:** *all columns except variables from exclusions*

**Derived Variables:**

**Assumptions and data limitations:** - the dataset is imbalanced.

## 4.0 Dependencies and Risks

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Likelihood (based on historical data)** | **Delay (based on historical data)** | **Impact** |
| As the rules and trends are changeable on the digital Ethereum platform, updated information of transactions is always needed to be able to detect the fraud transaction. | *Low* |  | *Once the analysis begins, we can detect the fraud transaction. However, we need more different fraud transactions to make the model familiar with any kind of fraud transaction.* |

## 5.0 Deliverable Timelines

|  |  |
| --- | --- |
|  | List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Item** | **Major Events / Milestones** | **Description** | **Scope** | **Days** | **Date** |
| 1. | Kick-off / Formal Request | *Meet with Advisor, choosing of dataset to be used, looking through the dataset* |  | 3 | *8th July 2022* |
| 2. | Analysis Plan | *Understanding the background of the related industry and completing the analysis plan* |  | 3 | *15th July 2022* |
| 3. | *EDA* | * Start working with python * plan and decide choosing right variables, analysis methods * data cleaning, handling duplicate rows, missing values, histograms, heatmap |  | 7 | *15Th July 2022* |
| 4. | Building models | * train data * build model * predict models * choose the best models |  | 10 | *25th July 2022* |
| 5. | Governance | risk and bias identification |  | 2 | *1st August 2022* |
| 6. | Documentation | Prepare the report for the project |  | 4 | *1st August 2022* |
| 7. | Internal team Presentation | Presentation |  | 5 | *8th August 2022* |
| 8. | Portfolio | Update the portfolio |  | 4 | *15th August 2022* |

# Works Cited

*CoinDesk*. (n.d.). Retrieved from coindesk.com: https://www.coindesk.com/price/ethereum/

DeMatteo, M. (2022, feb 10). *Opion*. Retrieved from coindesk.com: https://www.coindesk.com/tech/2022/02/10/the-top-ethereum-killers-compared/