Blockchain Fraud Prediction using Machine Learning

CSCI E-118 Project Proposal

by Kartik Srikumar

1) **Project Type:**

The project would fall under 'Blockchain Analysis' but will go beyond analysis. Some of the exploratory questions it will attempt to answer are:

- Are there a group of characteristics that often result in fraud over blockchain? What are they?
- Are some characteristics of an account or transaction more important in predicting fraud than others?

Additionally, the main intent of the project is to predict the occurrence of Fraud, given a set of features.

2) **Main Project Idea:**

To predict cases of blockchain fraud by leveraging machine learning. Specifically, I plan to use a dataset with multiple features such as value details of transactions sent by an address, time between transactions etc., along with the label 'Fraud' which will have binary values. I plan to use a Supervised Learning approach, comparing multiple classification models and selecting one or more of them, based on the trade-off between recall score and False Positive prediction.

3) **Potential Market:**

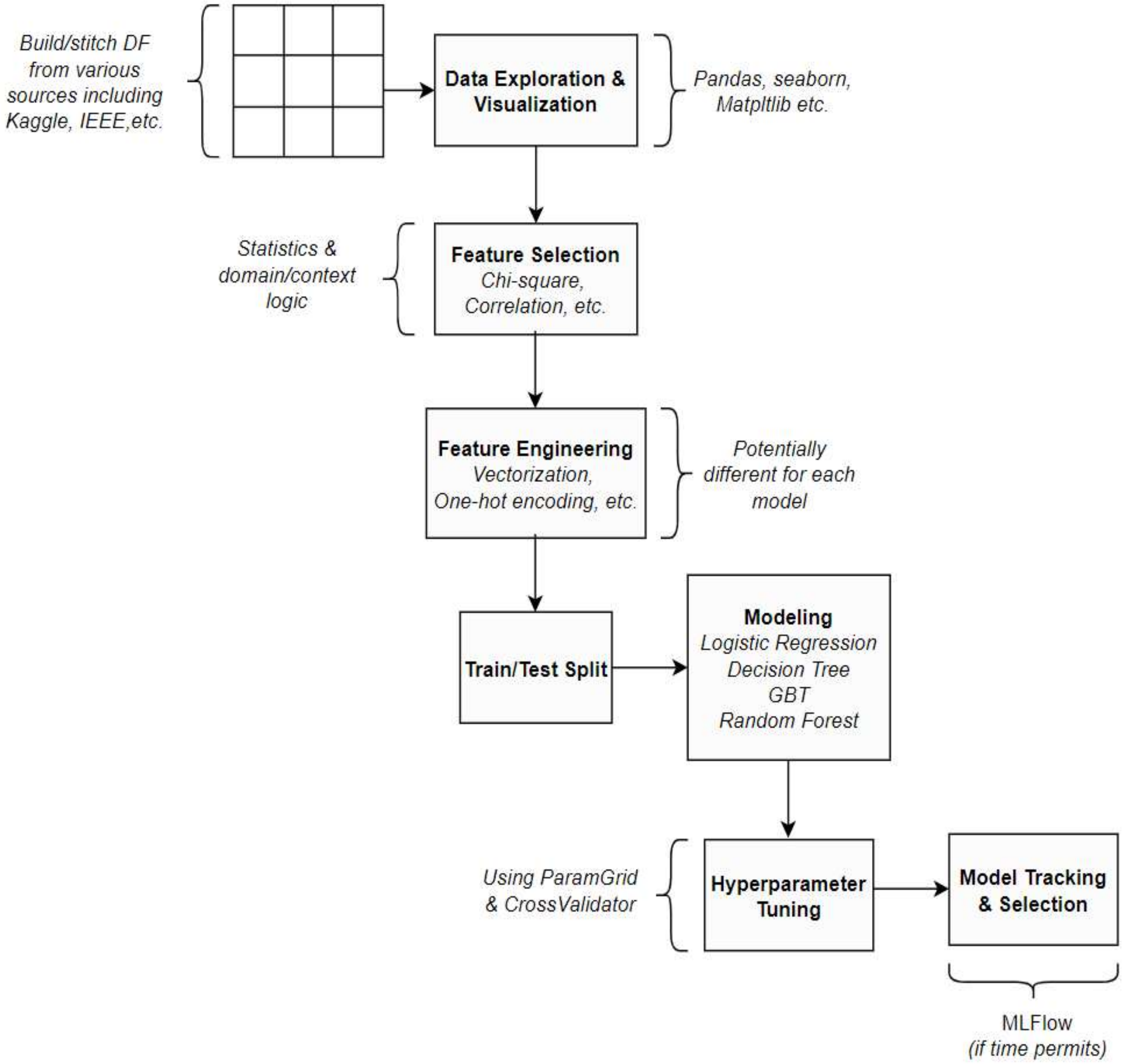
Since blockchain is decentralized, there would be no single authority that would want to implement this technology, but I envision it as being a utility that each node in the chain can optionally leverage to predict the outcome of transactions based on the model predictions. Mining pools are a great market opportunity.

4) **Architecture:**

The project will be implemented in python, while using several libraries including but not limited to sklearn, seaborn, MatplotLib, keras and pandas. The demo will be via a youtube video by leveraging a jupyter notebook. The notebook itself will have extensive documentation describing the workings of the system at each step.
Below are some of the tools I plan to use:

EDA & Visualization	Pandas, numpy, matplotlib.pyplot, seaborn
Feature Selection	sklearn.feature_selection.chi2, SelectKBest
Feature Engineering	sklearn.preprocessing.OneHotEncoder, CountVectorizer
Modeling & Hyperparameter Tuning	Cross-validation & Parameter Grid libraries/functions TBD
Model Tracking & Selection	MLFlow(*time permitting*), else manually built tracker using python.

Pipeline:


Build/stitch DF from various sources including Kaggle, IEEE,etc.
Data Exploration & Visualization
Pandas, seaborn, Matpltlib etc.
Statistics & domain/context logic
Feature Selection
Chi-square, Correlation, etc.
Feature Engineering
Vectorization, One-hot encoding, etc.
Potentially different for each model
Train/Test Split
Modeling
Logistic Regression
Decision Tree
GBT
Random Forest
Using ParamGrid & CrossValidator
Hyperparameter Tuning
Model Tracking & Selection
MLFlow
(if time permits)

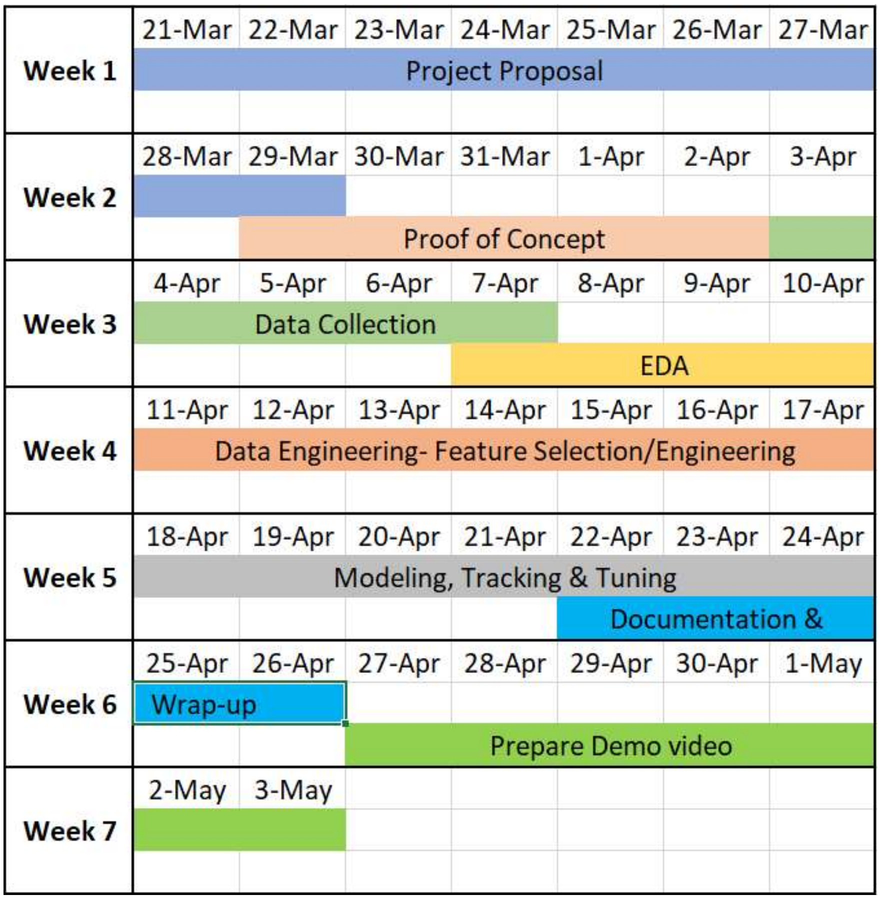
5) **Last Mile Analogue:** N/A

6) **Scaling Considerations:**
As time goes by, the training data will grow, which would mean opportunity to retrain the model with more data, potentially generalizing it more, but also having to deal with the challenge of storage and regular retraining and tuning. For the storage, we can use a distributed system like Hadoop or the cloud. For the continuous retraining and tuning, we can leverage a tool like MLFlow. Specifically, MLFlow's Tracking Server can be used to keep a track of the variations of models run along with metrics and artifacts. MLFlow Models can be used to load saved models.

7) **Fraud & Malicious Behavior:**
In some ways this project aims to curtail fraud and deter bad actors from successfully carrying out theft and other attacks. That said, it could be that in the future, bad actors are able to 'learn' about workarounds to the system to avoid getting caught by model predictions.

8) **Blockchain vs. Traditional Implementation:** N/A

9) **Project Timeline:**



Week 1	21-Mar	22-Mar	23-Mar	24-Mar	25-Mar	26-Mar	27-Mar
	Project Proposal						
Week 2	28-Mar	29-Mar	30-Mar	31-Mar	1-Apr	2-Apr	3-Apr
	🟦	🟦					
		Proof of Concept					🟩
Week 3	4-Apr	5-Apr	6-Apr	7-Apr	8-Apr	9-Apr	10-Apr
	Data Collection						
				EDA			
Week 4	11-Apr	12-Apr	13-Apr	14-Apr	15-Apr	16-Apr	17-Apr
	Data Engineering- Feature Selection/Engineering						
Week 5	18-Apr	19-Apr	20-Apr	21-Apr	22-Apr	23-Apr	24-Apr
	Modeling, Tracking & Tuning						
					Documentation &		
Week 6	25-Apr	26-Apr	27-Apr	28-Apr	29-Apr	30-Apr	1-May
	Wrap-up						
			Prepare Demo video				
Week 7	2-May	3-May					
	🟩	🟩					