Question:

Can we identify fraudulent transactions in financial data using machine learning?

Introduction:

In recent years, financial fraud has become a major problem for financial institutions around the world. Fraudulent activities like account takeovers, phishing attacks, and identity theft are causing huge financial losses. As a data analyst, I am interested in finding ways to prevent financial fraud using machine learning. In this proposal, I will identify a question and a data set that I will use to answer the question. I will also justify why this problem is important and why I think the data set will allow me to (begin to) answer the question.

Data Set:

The data set I will be using for this project is the "Synthetic Financial Dataset For Fraud Detection" available on Kaggle. The data set contains over 6 million rows of transactional data with 11 features including step, type, amount, nameOrig, oldbalanceOrg, newbalanceOrig, nameDest, oldbalanceDest, newbalanceDest, isFraud, and isFlaggedFraud.

Importance:

The importance of this project lies in the prevention of financial fraud. Financial institutions can use machine learning algorithms to detect fraudulent transactions and prevent losses. By identifying fraudulent transactions early, financial institutions can protect their customers' accounts and prevent further damage.

Data Set Suitability:

The data set contains a large amount of transactional data, with various features that can be used for machine learning models. The data set also contains a good mix of legitimate and fraudulent transactions, which will allow us to train and test our models.

Possible Process:

The proposed project will involve a thorough data analysis process that includes feature selection, feature correlation analysis, data visualization, data cleaning, regression with feature engineering, and a comparison of different machine learning models. The aim is to identify the most important features for predicting fraud and to determine which machine learning model performs best on the given dataset. This process will help to build an accurate predictive model to prevent fraudulent activities in the financial system.

Conclusion:

In conclusion, financial fraud is a major problem for financial institutions, and machine learning can be used to detect fraudulent transactions and prevent losses. The "Synthetic Financial Dataset For Fraud Detection" is a suitable data set for this project, and by identifying fraudulent transactions early, financial institutions can protect their customers' accounts and prevent further damage.